Some Problems in Testing for Modality

Matthew Gerard York

A thesis submitted for the degree of Doctor of Philosophy of
The Australian National University

March 1999

Declaration

Unless otherwise specified in the text, this thesis described my own work, supervised by Professor P.G. Hall.

Matthew Gerard York

Acknowledgements

Firstly, I would like to thank my supervisor, Peter Hall, for his generosity, kindness, guidance and encouragement. His assistance has been invaluable and greatly appreciated. The financial assistance of an Australian Postgraduate Award and an ANU Supplementary Scholarship is gratefully acknowledged.

Special thanks go to all my friends and family, especially my mother, for their constant support during the preparation of this thesis.

I would like to thank all the staff, students and visitors in the Statistical Science Program of the Centre for Mathematics and Its Applications for creating such a pleasant and stimulating research environment.

Abstract

The topic of this thesis is testing for modality. Chapter 1 provides an introduction to the thesis and gives a survey of the existing literature on the problem of testing for modality. One of the most widely used techniques for mode testing is Silverman's critical bandwidth test. This bootstrap test is known to be conservative and in Chapter 2 we propose a means of calibrating the bandwidth test to improve its level accuracy. We describe the general properties of our calibrated form of the test and develop theory describing the test. We also address the problem of testing for the number of modes of a density in a compact interval, rather than over the whole real line. In Chapter 3 we quantify the conservatism of Silverman's test and we study the numerical performance of our proposed test in an extensive simulation study. The test is applied to real data sets in Chapter 4. Chapter 5 deals with the related problem of estimating the components in a mixture of smooth regression curves. We develop an algorithm which employs local linear methods to estimate the curves nonparametrically. We propose bandwidth selectors based on plug-in ideas. We use our calibrated test for modality to determine the number of components in the mixture.

Contents

Declaration i

Acknowledgements ii

Abstract iii

1 Introduction 1

1.1 Introduction 1
1.2 Bump hunting 3
1.3 Silverman's critical bandwidth test 7
1.4 The DIP test 11
1.5 Excess mass estimates and tests 12
1.6 The mode tree and mode existence test 17
1.7 Mode testing in higher dimensions 19
1.8 Finite mixture distributions 22

2 Calibrating Silverman's test 27

2.1 Introduction 27
2.2 The bootstrap test 28
2.3 Consistency of the test 29
2.4 Spurious modes and compact intervals 30
2.5 Testing for j modes 32
2.6 Theoretical results 33
2.6.1 Separation of the modes of $\hat{f}_h$ 33
2.6.2 Distribution of $\hat{h}_{\text{crit}}$ under H_0 33
2.6.3 Bootstrap distribution of $\hat{h}^*_{\text{crit}}/\hat{h}_{\text{crit}}$ under H_0 34
2.6.4 Distribution of $\hat{h}_{\text{crit}}$ and $\hat{h}^*_{\text{crit}}/\hat{h}_{\text{crit}}$ under H_1 35

2.6.5 Alternative regularity conditions 36
2.7 Technical arguments 36
2.7.1 Proof of Theorem 2.1 36
2.7.2 Proof of Theorem 2.2 37
2.7.3 Proof of Theorem 2.3 40
2.7.4 Proof of Theorem 2.4 43
2.7.5 Proof of Theorem 2.5 44

3 Numerical properties **48**
3.1 Introduction 48
3.2 Calibration 48
3.3 Large-sample behaviour of the test 55
3.4 Behaviour of the test on Normally distributed samples of finite size . 60
3.4.1 Effect of sample size 60
3.4.2 Effect of the choice of the interval $\mathcal{I}$ 62
3.5 Corrections to reduce conservatism 66
3.6 Performance of the corrections 70
3.7 Calibrating the test for finite sample sizes 77
3.8 Power of the test 82

4 Examples **98**
4.1 Introduction 98
4.2 Chondrite data 100
4.3 Buffalo snowfall data 103
4.4 Swiss bank notes data 106
4.5 Stamp data 111
4.6 Galaxy velocity data 117

5 Mixture data and curve estimation **122**
5.1 Introduction 122
5.2 Estimating two means in a mixture 122
5.3 Nonparametric curve estimation with mixture data 127
5.4 Bandwidth selection 132
5.5 Assessing the number of components in a mixture of smooth curves . 140
5.6 Practical performance 142

References **153**

Chapter 1

Introduction

1.1 Introduction

This thesis is primarily concerned with assessing the modality of a population. Since most commonly considered density functions are unimodal, the presence of more than one mode is generally interpreted as a sign that the data are clustered. By clustered, we mean that the population from which the data are drawn is not homogeneous but instead is made up of a number of more homogeneous subpopulations. We are interested in formally testing whether structure that appears in data sets reflects true features of the underlying density function. We focus our attention on the important case of testing whether a populations is homogeneous, with a unimodal density function, or whether it is multimodal.

The assessment of modality plays an important role in many applications, ranging from the study of fundamental scientific theories about the structure of the universe and the existence of new elementary particles to understanding curious features of collectables. For example, Roeder (1990) examined the distribution of the velocities of a sample of galaxies. The velocity of a galaxy is an indication of its distance from the Earth. Astronomers predicted that gravitational attraction would lead to some clustering. Roeder's analysis of these data indicates that the data are highly multimodal. This suggests the existence of superclusters of galaxies, surrounded by large voids. The forces causing this large-scale clustering cannot be entirely explained by existing theory. Hence in this case the detection of multimodality has confirmed the need for a reassessment of current theories.

Good and Gaskins (1980) investigated the existence of modes and bumps in

mass spectra in high-energy physics scattering experiments. The detection of modes and bumps can give evidence of the existence of new elementary particles. The discovery of bimodality in lipid data by Scott *et al.* (1980) has led to clarification of the roles of certain risk factors in coronary artery disease (Scott, 1980). Izenman and Sommer (1988) examined the distribution of the thickness of stamps from a nineteenth century Mexican stamp issue and deduced from the highly multimodal structure of the density function that the stamp issue had been printed on a greater number of different types of paper than had previously been thought by collectors.

Traditionally, mixture distributions have been used to model multimodal distributions. The assessment of modality is often an important part of the analysis of mixture models since it gives an indication of the number of components in the mixture. Knowing the likely number of modes of a density can also be very useful in nonparametric density estimation. Bump hunting and mode testing reveal important features in a population and we would normally want to choose the amount of smoothing of a nonparametric density estimator so that the estimate exhibits these features. Cuevas and Gonzales-Manteiga (1991) describe methods of bandwidth selection for kernel density estimators that match the number of modes of the estimated and true densities.

In the remainder of this chapter we shall give a survey of existing methods for bump hunting and assessing the modality of a population. We shall mainly restrict our attention to univariate populations. In Section 1.7 we shall look at techniques for generalising these univariate approaches to higher dimensional problems. Finite mixture distributions and their connections with assessing multimodality will be described in Section 1.8.

Silverman's (1981, 1983) critical bandwidth test is the most popular nonparametric technique for testing for multimodality but it is well known that the test is very conservative, in the sense that its actual level tends to be much less than its nominal one. In Chapter 2 we propose a means of calibrating the test to improve its level accuracy and increase its power. We also address the problem of testing for the number of modes of a density in a compact interval. In the past only the case of testing for the number of modes on the whole line has received much attention. In theoretical studies this has restricted attention to densities with compact support and in practice it can lead to difficulties with spurious modes arising from outlying data. We describe the general properties of our new forms of the test and

develop theory that shows that our calibration method produces a test that has asymptotically correct level.

Chapter 3 focuses on numerical aspects of the test. We quantify the conservatism of Silverman's test and we calculate the corrections that are required calibrate the test. Through an extensive simulation study we aim to obtain a greater understanding of the test and to assess the performance of our methods. We also examine the improvement in the power of the test that our calibration methods deliver.

The test is applied to a number of real data sets in Chapter 4 in order to illustrate the effectiveness of our methods for data analysis.

In Chapter 5 we consider the closely related problem of estimating the components in a mixture of two smooth regression curves. We develop an algorithm for estimating smooth curves from mixture data and propose a technique for determining the presence of a mixture for regression data.

1.2 Bump hunting

Testing for the existence of modes can be thought of as a subproblem in a wider area of interest, that of bump hunting. In one dimension we shall define a mode of a density f to be a point where $f'(x) = 0$ and $f''(x) < 0$. A bump will be an interval over which $f''(x) < 0$. Hence a bump does not necessarily contain a mode but a mode is always located on a bump. Cox (1966) proposed a method for testing the existence of bumps on a histogram and Good and Gaskins (1980) developed a procedure for testing for the existence of a bump using smooth nonparametric density estimates.

Good and Gaskins' procedure is based on estimating the density, then locally smoothing away each bump, one at a time, in order to assess the significance of each bump. Good and Gaskins (1980) used the method of maximum penalised likelihood (Good and Gaskins, 1971) to estimate the density. For an independent and identically distributed sample $X_1, \ldots, X_n$, the maximum penalised likelihood estimator $\hat{f}$ is the function f that maximises the penalised likelihood, defined by

$$\tilde{L}(f) = \prod_{i=1}^{n} f(X_i)\, e^{-\Phi(f)}, \tag{1.1}$$

or equivalently, maximises the penalised log-likelihood,

$$\omega = \omega(f) = L(f) - \Phi(f) = \sum_{i=1}^{n} \log f(X_i) - \Phi(f), \tag{1.2}$$

where L denotes the log-likelihood and $\Phi(f)$ is a roughness penalty. The first term, L, is a measure of the goodness-of-fit to the data and the penalty measures the smoothness of f. Without the roughness penalty the maximiser $\hat{f}$ would be a collection of spikes at each of the data points. The roughness penalty imposes a degree of smoothness on the density estimate. Hence, maximising penalised likelihood is a trade-off between smoothness and goodness-of-fit. For more details on maximum penalised likelihood as a technique for probability density estimation, see Chapter 4 of Tapia and Thompson (1978) and Section 5.4 of Silverman (1986).

Good and Gaskins (1980) use the roughness penalty

$$\Phi(f) = \beta \int \{\gamma''(x)\}^2 \, dx \,,$$

where $\gamma(x) = \{f(x)\}^{1/2}$ and β is a positive smoothing parameter that needs to be determined. Large values of β will result in a density estimate with increased smoothness at the expense of goodness-of-fit to the data and smaller values of β will yield bumpier estimates that provide a closer fit to the data.

The roughness penalty has a simple Bayesian interpretation which is useful for assessing the significance of a bump. We can regard $e^{-\Phi(f)}$ in equation (1.1) as being proportional to an improper prior density over the space of smooth functions f, so the smoothing parameter β is a hyperparameter, a parameter of the prior. The penalised log-likelihood $\tilde{L}$ can then be shown to correspond, up to a constant, to the logarithm of the posterior density. Thus the maximised penalised likelihood estimator $\hat{f}$ is the mode of the posterior density over the space of smooth functions (Silverman, 1986, p. 119).

Good and Gaskins (1971, 1980) employ orthogonal series to estimate $\gamma(x)$. This method makes use of the expansion

$$\gamma(x) = \sum_{m=0}^{r-1} \gamma_m \, \phi_m(x) \,, \tag{1.3}$$

where $\phi_0(x)$, $\phi_1(x), \ldots, \phi_{r-1}(x)$ are normal orthogonal functions and $\gamma_1, \ldots, \gamma_{r-1}$ are the corresponding coefficients. Good and Gaskins (1980) give a number of reasons for preferring the use of the Fourier series and suggest a way of determining

r. In theory r is infinite but in practice it needs to be terminated in the interests of computational efficiency. The series is not being terminated to control the smoothness of $\hat{f}$, this is entirely the work of the smoothing parameter, β.

To implement this method it is necessary to select the value of the smoothing parameter β. This could be done using cross-validation but Good and Gaskins opt for a technique based on goodness-of-fit statistics, in particular the χ^2 and the Kolmogorov-Smirnov statistics. Let S be a goodness-of-fit statistic and let P_S be the probability, conditional on $\hat{f}$ being the true density, that S is less than the observed value. If β is taken too large, the density $\hat{f}$ will be oversmooth and P_S will be close to 1 since there will be a poor goodness-of-fit, that is S will be large. For too small a value of β the estimate will resemble the observations too much and P_S will be close to zero. Good and Gaskins argue that the optimal value of P_S is $1/2$ as it treats smoothness and goodness-of-fit symmetrically.

Since χ^2 statistics can be quite sensitive to the choice of class boundaries on which they are based, it is recommended to compute several χ^2 statistics based on different, co-prime numbers of class intervals. Good and Gaskins (1980) suggest a way of combining the probabilities corresponding to the Kolmogorov-Smirnov and χ^2 statistics to obtain a single probability.

Once the smoothing parameter is determined the maximum penalised likelihood estimate $\hat{f}$ of the density can be constructed and the bumps that appear in the density can be evaluated. Good and Gaskins (1980) give a Bayesian technique for evaluating each bump one at a time. They compare a hypothesis $f_1(x)$ that shows the bump with a similar hypothesis $f_2(x)$ but with just that bump removed. If there exists theory for predicting the existence of a bump in a specific location then it would be reasonable to use

$$L(f_1) - L(f_2)$$

as the weight of evidence in favour of f_1 against f_2. If there is no theoretical reason for believing there is a bump in a specific location then it is more appropriate to use

$$\omega(f_1) - \omega(f_2),$$

the difference in the penalised log-likelihoods for the two hypotheses. The result will then be the final posterior log-odds in favour of the existence of the bump because the scores incorporate the prior that we have selected by fixing the hyperparameter β. Good and Gaskins (1980) notice in their example that the evaluation of bumps is not particularly sensitive to the chosen value of the smoothing parameter.

The density f_2, which is exactly the same as the density f_1 but with the bump removed, is calculated objectively using an iterative procedure. Since all smoothed density estimates are biased downwards at bumps the bump can be removed by repeated smoothing. This is done by adjusting the observations in the vicinity of the bump to make the observed data agree with the density estimate and then renormalising the estimate. Then another smoothed density is fitted to the adjusted data and the data are adjusted again. This smoothing process is repeated until convergence is achieved. Good and Gaskins report that convergence normally occurs in fewer than a dozen iterations, except in the case of extremely large bumps that are obviously true features of the underlying population.

The strength of Good and Gaskin's approach to bump hunting is the ease with which the significance of each bump can be computed, once the density estimate has been constructed. By making use of the Bayesian interpretation of maximum penalised likelihood, the log-odds in favour of a bump are readily calculated. The other techniques that we consider in this chapter require the use of computationally intensive resampling methods, which can be very conservative, or even more conservative methods that rely on the use of a reference null distribution to assess the significance of bumps and modes.

However the weakness of this approach to bump hunting is the number of choices that need to be made by the user. There is the choice of an orthogonal series and a truncation point r to be made in (1.3) and, most importantly, the value of the smoothing parameter β needs to be determined. While Good and Gaskins (1980) report that for their examples the odds in favour of a bump did not change much within a narrow range of choices of β, it is not clear how sensitive the odds are for less precise choices of β. We see in the next section that the modality of a kernel density estimator is entirely dependent on choice of the smoothing parameter, and maximum penalised likelihood estimators would be expected to behave similarly. Silverman (1981) exploits this dependence of the modality of a kernel density estimator on its bandwidth to develop a test for the number of modes in a population. This approach leads to a natural and automatic choice for the smoothing parameter.

Good and Gaskins' procedure does not appear to have been studied in any depth either theoretically or practically. Consequently it is unclear how it performs in practice, in particular it is not known how the performance of the test depends on choice of smoothing parameter nor how powerful it is.

1.3 Silverman's critical bandwidth test

Silverman (1981, 1983) developed a method for investigating the number of modes in a population. Like the bump hunting technique of Good and Gaskins (1980) it is based on density estimates but it differs in that this test is based on the kernel density estimator and the amount of smoothing is chosen automatically in a natural way. This makes the test both easy to implement and intuitively appealing. These two factors have combined to contribute to the popularity of this test.

Given an independent sample $\mathcal{X} = \{X_1, \ldots, X_n\}$ from a distribution with unknown density f, we wish to test the null hypothesis H_0 that f has precisely j modes against the alternative H_1 that f has more than j modes. We begin by constructing the kernel density estimator

$$\hat{f}_h(x) = (nh)^{-1} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) ,$$

where h is a bandwidth and K is a kernel function. Throughout this thesis we shall take K to be the standard Normal density function. This choice has strong theoretical advantages as well as being computationally attractive; see Silverman (1982).

The bandwidth h controls the amount the data are smoothed to obtain the kernel density estimator. For example, for data from a distribution whose density has more than j modes a large value of h will be required to yield a density estimate with j modes. This suggests using the smallest bandwidth that yields a density estimate with j modes as the test statistic for H_0. Define the critical bandwidth to be

$$\hat{h}_{\text{crit},j} = \inf \left\{ h : \hat{f}_h \text{ has at most } j \text{ modes} \right\} .$$

Large values of the critical bandwidth will reject the null hypothesis.

Silverman (1981) showed that when the Gaussian kernel is used the number of modes of a kernel density estimate is a non-increasing function of the bandwidth h. This result simplifies the evaluation of $\hat{h}_{\text{crit},j}$ and adds to the intuitive appeal of using $\hat{h}_{\text{crit},j}$ as a test statistic. This result does not hold for general kernels, even unimodal ones; see for example Hart (1984). In fact, this property is unique to the Normal kernel among all commonly considered kernels, including the Uniform, Epanechnikov, Biweight and Triweight kernels. See Marron and Nolan (1989) for a description of these kernels. Silverman's proof relies on two special properties of the

Gaussian kernel. These are the total positivity (see Karlin, 1968) of the Gaussian density and the fact that the convolution of a Normal kernel density estimator and a Normal kernel produces another Normal kernel density estimator.

Silverman (1983) and Mammen, Marron and Fisher (1992) have extensively studied the theoretical properties of $\hat{h}_{\text{crit},j}$. Their theoretical results show that under H_0, $\hat{h}_{\text{crit},j}$ tends to zero and under H_1, $\hat{h}_{\text{crit},j}$ is bounded away from zero. Hence, using $\hat{h}_{\text{crit},j}$ as a test statistic, where large values of $\hat{h}_{\text{crit},j}$ lead to H_0 being rejected, will produce a consistent test. Stating these results more precisely, Mammen, Marron and Fisher (1992) showed that if H_0 is true, and under appropriate conditions on f (see the statement of Theorem 2.2 of this thesis for the precise conditions),

$$\lim_{C_1 \downarrow 0, C_2 \uparrow \infty} \liminf_{n \to \infty} P\left(C_1 n^{-1/5} < \hat{h}_{\text{crit},j} < C_2 n^{-1/5}\right) = 1 .$$

Silverman (1983) proved that if H_1 is true then there exists a constant $c > 0$, depending on f and j, such that

$$\liminf_{n \to \infty} P\left(\hat{h}_{\text{crit},j} > c\right) = 1 .$$

To implement the test we need to have a method of determining when $\hat{h}_{\text{crit},j}$ is too large to be consistent with the null hypothesis. Silverman (1981) suggested the use of a smoothed bootstrap technique for assessing significance. Let $\hat{f}_{\text{crit}}$ denote the version of $\hat{f}_h$ obtained by putting $h = \hat{h}_{\text{crit},j}$. Conditional on $\mathcal{X}$, let $X_1^*, \ldots, X_n^*$ be a resample drawn from the distribution with density $\hat{f}_{\text{crit}}$, and put

$$\hat{f}_h^*(x) = (nh)^{-1} \sum_{i=1}^{n} K\left(\frac{x - X_i^*}{h}\right) .$$

Let $\hat{h}^*_{\text{crit},j}$ denote the version of $\hat{h}_{\text{crit},j}$ in this setting, that is, the infimum of all bandwidths such that $\hat{f}^*_{\text{crit}}$ has at most j modes. Silverman (1981) used the bootstrap distribution function

$$P\left(\hat{h}^*_{\text{crit},j} \leq \hat{h}_{\text{crit},j} \,\middle|\, \mathcal{X}\right) \tag{1.4}$$

to estimate the distribution of $\hat{h}_{\text{crit},j}$, and one minus the probability (1.4) was used as a p-value.

The distribution with density $\hat{f}_{\text{crit}}$ is chosen as the distribution from which to resample since it is the density estimate with j modes (that is, it is consistent with

H_0) that is closest to the data. By closest we mean that it uses the least amount of smoothing among all estimates that satisfy H_0. Since a kernel density estimate is a convolution of the empirical distribution function with the kernel function we can generate independent realisations X^* from $\hat{f}_{\text{crit}}$ by

$$X_i^* = X_{I(i)} + \hat{h}_{\text{crit},j}\epsilon_i\,, \tag{1.5}$$

where $X_{I(i)}$ are sampled uniformly, with replacement, from $\mathcal{X}$, and ϵ_i are independent standard Normal random variates.

Silverman (1986, Section 6.4.1) recommends that rather than resampling from $\hat{f}_{\text{crit}}$ we should resample from a rescaled version of it that has the same first two moments as the original data $\mathcal{X}$. It can easily be seen that resamples generated using the scheme (1.5) will have a variance of $\sigma^2 + \hat{h}^2_{\text{crit},j}$, where σ^2 is the variance of the original data. It is recommended to use the refinement, suggested by Efron (1979),

$$X_i^* = \bar{X} + (X_{I(i)} - \bar{X} + \hat{h}_{\text{crit},j}\epsilon_i)/(1 + \hat{h}^2_{\text{crit},j}/\hat{\sigma}^2)^{1/2}\,, \tag{1.6}$$

where $\bar{X}$ and $\hat{\sigma}^2$ are the sample mean and variance of $\mathcal{X}$. It will be seen in Chapter 3 that this correction greatly improves performance of the test for finite sized samples.

Silverman (1983) observed that by resampling from $\hat{f}_{\text{crit}}$, a density on the boundary of H_0 and H_1, the bootstrap probability (1.4) will provide a conservative assessment of the significance of H_1. In fact this method is extremely conservative. The simulations of Mammen, Marron and Fisher (1992) and Fisher, Mammen and Marron (1994) provide numerical evidence of this and Mammen *et al.* present a very convincing heuristic argument, along with theoretical results on $\hat{h}^*_{\text{crit},j}$, to explain this conservatism. Chapters 2 and 3 of this thesis examine the extent of this conservatism and suggest ways of improving the accuracy of the test.

Alternatives to the bootstrap for assessing the significance of the test, for the case of $j = 1$, are assessing $\hat{h}_{\text{crit},j}$ against a standard family of unimodal distributions (Silverman, 1986, p. 140) or approximations based on the asymptotic distribution of $\hat{h}_{\text{crit},j}$ (Hall and Wood, 1996). Hall and Wood (1996) develop conservative Normal approximations to the asymptotic distribution of $\hat{h}_{\text{crit},j}$ and argue that these methods are more accurate than the bootstrap method described above since this simple bootstrap method does not produce a consistent estimate of the limiting distribution of $\hat{h}_{\text{crit},j}$. In Chapter 2 we develop an accurate bootstrap method. Hall and Wood

also investigate the related problems of testing for the number of shoulder points of a density (points where $f'(x) = f''(x) = 0$) and the number of points of inflexion (points where $f''(x) = 0$).

Over the past fifteen years the critical bandwidth test has become the most popular method for determining the number of modes in a population. In many ways this popularity is due to the test being based on the kernel density estimator. Since the kernel density estimator is the most widely used and understood nonparametric density estimator, after the histogram, the ideas on which the test is based are generally well understood and hence the test is an intuitively appealing one. In addition, the most important and often most troublesome feature of kernel estimation is the choice of the bandwidth and this problem has been widely studied without an entirely satisfactory solution being found. Jones, Marron and Sheather (1996) provide a recent survey of bandwidth selection techniques for density estimation. However for the bandwidth test the bandwidth is chosen automatically in a very natural manner, and this makes the test extremely easy to implement.

However, a test that depends on a density estimator will inherit any undesirable properties that estimator might have. In the context of mode testing, the most serious of these properties of a kernel density estimator with a global bandwidth is that these estimators will often have spurious modes in their tails, caused by the sparsity of data towards the edge of the sample. This can result in the test falsely detecting non-existent modes in the tails of a distribution. Another problem is that the test has troubles with densities that are difficult to estimate using a kernel estimator with a global bandwidth. For example, if a density has a number of modes of widely varying shapes then the test has difficulties detecting the correct number of modes (Minnotte, 1997).

In Chapter 2 we address the situation of testing for the number of modes over a compact interval, instead of the usual practice of testing for modes over the whole real line. This approach allows us to avoid the problem of detecting spurious modes in the tail of a distribution and it can alleviate other weaknesses of using a global bandwidth by allowing the use of different bandwidths over different intervals.

1.4 The DIP test

Hartigan and Hartigan (1985) proposed the *DIP statistic* as a measure of the multimodality of a sample. It is the maximum difference between the empirical distribution function and the unimodal density that minimises that distance. This statistic has the advantage of being based on the empirical distribution function and does not require the explicit estimation of a density, unlike the procedures of Good and Gaskins (1980) and Silverman (1981).

The DIP is defined by

$$D(F) = \min_{G \in \mathcal{U}} \max_{x} |F(x) - G(x)|,$$

where $\mathcal{U}$ is the class of unimodal functions. The DIP can be used to measure departures from unimodality since $D(F) = 0$ for $F \in \mathcal{U}$ and $D(F) > 0$ for $F \notin \mathcal{U}$. The DIP test of unimodality of a sample of size n is based on calculating the DIP for the empirical distribution function $\widehat{F}_n$. Define

$$\rho(\widehat{F}_n, F) = \max_{x} |\widehat{F}_n(x) - F(x)|$$

and observe that the following two inequalities

$$D(\widehat{F}_n) \leq D(F) + \rho(\widehat{F}_n, F)$$

and

$$D(F) \leq D(\widehat{F}_n) + \rho(\widehat{F}_n, F)$$

hold. The Glivenko-Cantelli theorem states that $\rho(\widehat{F}_n, F) \to 0$ almost surely, which implies that $D(\widehat{F}_n) \to D(F)$ almost surely. Asymptotically $D(\widehat{F}_n)$ will be zero for samples generated from unimodal distributions and will be positive otherwise. Therefore a test based on the DIP will asymptotically distinguish any unimodal F from any multimodal F.

Hartigan and Hartigan (1985) developed an algorithm for computing $D(\widehat{F}_n)$ from a sample of size n. They show that the statistic may be computed in order n operations. A FORTRAN implementation of the algorithm has been published by P.M. Hartigan (1985). A modal interval is produced as an outcome of the DIP calculation and this estimate of the mode has the fairly rare advantage of not requiring the specification of a kernel bandwidth or any other smoothing parameter.

In testing the null hypothesis that a distribution is unimodal versus the alternative that it is multimodal a method is needed to calculate the significance of the computed DIP. Hartigan and Hartigan (1985) suggest that this could be done by following Silverman (1981) and resampling from the unimodal distribution that is closest to the data; in this case closest means the distribution that minimises the DIP. However they recommend the simpler, but more conservative, option of using the Uniform distribution as a unimodal, null distribution.

The Uniform is chosen as the reference distribution since Hartigan and Hartigan conjecture that the DIP statistic $D(\widehat{F}_n)$ is asymptotically largest for the Uniform distribution amongst all unimodal distributions. Using the "least-favourable" unimodal distribution to assess significance will result in a conservative test. Unfortunately the DIP is not stochastically larger for the Uniform than for any other unimodal distribution, for all sample sizes n, but they prove two asymptotic results that convincingly support their conjecture.

The first is that, for a sample from the Uniform on $(0,1)$, $\sqrt{n}\,D(\widehat{F}_n) \to D(B)$ in distribution as $n \to \infty$, where B is a standard Brownian bridge. The second is that, for unimodal distributions whose densities decrease exponentially away from the mode, $\sqrt{n}\,D(\widehat{F}_n) \to 0$ in probability. These results support the conjecture since they show that $\sqrt{n}\,D(\widehat{F}_n)$ is asymptotically positive for the Uniform but is asymptotically zero for a wide class of unimodal distributions.

Hartigan and Hartigan (1985) give the commonly used percentage points of the DIP statistics for samples of a range of sizes, drawn from the Uniform distribution. They also perform some power calculations. Sommer and McNamara (1987) examine the power of the test when the underlying distribution is a known mixture of two Normal distributions.

1.5 Excess mass estimates and tests

Müller and Sawitzki (1991) propose *excess mass estimates* as a method for analysing the modality of a distribution. They prefer to think of modes in the statistical terms of being a region where an excess of probability mass is concentrated rather than in the analytical terms of being a local maximum of the density. Their approach is based directly on the empirical distribution function which means that it does not require the explicit estimation of a density and it allows investigation of the number

of modes to be separated from questions concerning their location.

For a distribution with density f the excess mass functional is

$$E(\lambda) = \int \left[f(x) - \lambda\right]_+ dx\,,$$

which is the amount of probability mass concentrated above a certain level $\lambda > 0$. It can be considered to be a measure of the "distinctiveness" of a mode.

$E(\lambda)$ is a sum of contributions $E_C(\lambda) \equiv \int_C [f(x) - \lambda]\, dx$, coming from the connected sets C, where $C \subset \{x : f(x) \geq \lambda\}$. The connected components of $\{x : f(x) \geq \lambda\}$ are known as λ-clusters and as λ increases the λ-clusters concentrate on modes. For any λ, a distribution with m modes has at most m λ-clusters. The excess mass $E_m(\lambda)$ for m modes is defined as

$$E_m(\lambda) = \sup_{C_1,\ldots,C_m} \sum_{j=1}^{m} \int_{C_j(\lambda)} \left[f(x) - \lambda\right] dx\,, \tag{1.7}$$

where the supremum is taken over all sequences $\{C_1, \ldots, C_m\}$ of disjoint connected intervals. Equation (1.7) can be written as

$$E_m(\lambda) = \sup_{C_1,\ldots,C_m} \sum_{j=1}^{m} \left\{F(C_j) - \lambda \|C_j\|\right\}, \tag{1.8}$$

where $F(C)$ is the F-measure of C and $\|C\|$ is the length of the interval C.

Given a sample $\mathcal{X} = \{X_1, \ldots, X_n\}$ drawn from the distribution F we substitute the empirical distribution function $\widehat{F}_n$ for F in (1.8) to obtain the estimator

$$E_{nm}(\lambda) = \sup_{C_1,\ldots,C_m} \sum_{j=1}^{m} \left\{\widehat{F}_n(C_j) - \lambda \|C_j\|\right\}. \tag{1.9}$$

To test the null hypothesis that f has $m-1$ modes against the alternative that it has m modes Müller and Sawitzki (1991) suggest that a large difference

$$D_{nm}(\lambda) = E_{nm}(\lambda) - E_{n,m-1}(\lambda)\,,$$

for some λ indicates a violation of the null hypothesis. Hence they recommend the use of

$$\Delta_{nm} = \sup_{\lambda > 0} D_{nm}(\lambda)$$

as the test statistic for the *excess difference test*, with large values of Δ_{nm} leading to the rejection of the null hypothesis.

Müller and Sawitzki (1991) present an algorithm for finding the components C_j that minimise the sum in equation (1.9), which they call the empirical λ-clusters. The algorithm also computes the excess mass estimate $E_{nm}(\lambda)$. In the absence of flat parts of f the empirical λ-clusters consistently estimate the real λ-clusters. The empirical λ-clusters show the location of mass concentration and plotting them against λ can be used as a data analysis tool. This idea is similar to mode tree of Minnotte and Scott (1993); see Section 1.6.

It is interesting to observe that in the case of testing for two modes versus one mode the excess mass test is identical to Hartigan and Hartigan's (1985) DIP test. The excess mass statistic Δ_{n2} is equal to twice the DIP statistic. Hence the following discussion of ways of developing a method for accurately assessing the significance of the excess mass statistic obviously also applies to the DIP statistic.

In the previous section we saw that the significance of a calculated DIP statistic is usually determined by reference to the distribution of the DIP calculated for samples drawn from a Uniform distribution. This choice of the Uniform as a reference distribution was justified by theoretical results which showed that the DIP was asymptotically larger for the Uniform samples than for samples drawn from unimodal distributions whose densities decrease exponentially away from the mode.

Similarly for the excess mass test, Müller and Sawitzki have shown that for samples drawn from a Uniform distribution Δ_{nm} is of exact order $O_p(n^{-1/2})$, and Cheng and Hall (1998) have shown that Δ_{nm} is of exact order $O_p(n^{-3/5})$ for distributions without any flat parts (see Cheng and Hall (1998) for the precise regularity conditions). Therefore the use of the Uniform as a reference distribution will produce an asymptotically conservative test. In fact, it will be extremely conservative. Since Δ_{nm} is of a larger exact asymptotic order for the Uniform than for distributions without any flat parts, the use of the Uniform as a reference distribution will result in a test with an asymptotic level of zero for each non-zero value of the nominal level (Cheng and Hall, 1998). The conservatism of the test results in it having rather low power.

Cheng and Hall (1997, 1998) develop two forms of the test both of which produce a test that is asymptotically accurate. They concentrate on the case of testing the null hypothesis of unimodality versus the alternative of multimodality. Their approaches exploit the fact that the limiting distribution of Δ_{nm}, under the null hypothesis that f is unimodal, depends on unknowns only through a constant, which

may be estimated. They prove that $n^{3/5}\,\Delta_{n2}$ converges in distribution to cZ as $n \to \infty$, where $c = \{f(x_0)^3/|f''(x_0)|\}^{1/5}$, x_0 is the location of the unique mode of f and Z is a random variable whose distribution does not depend on f.

Their first method requires resampling from a distribution F^0, for which the properties of Δ_{n2} are similar to those under F if the null hypothesis is true. Since the limiting distribution of Δ_{n2} depends only on f through the constant c, the only requirement on F^0 is that it has a unimodal density f^0 with the same value of c as f. In their implementation Cheng and Hall use kernel methods to estimate f and f''. Notice that the estimates $\hat{f}$ and $\widehat{f''}$ require different amounts of smoothing. They use the respective asymptotically optimal global bandwidths for a Normal $N(0, s^2)$ distribution, where s^2 is the sample variance. Once c has been estimated by $\hat{c} = \{\hat{f}(\hat{x}_0)/\widehat{f''}(\hat{x}_o)\}^{1/5}$, where $\hat{x}_0$ is the location of the largest mode of $\hat{f}$, they choose f^0 to be a density from a family of Beta, Normal and Student-t densities with a value of c equal to $\hat{c}$.

The procedure is to draw a resample $\mathcal{X}^* = \{X_1^*, \ldots, X_n^*\}$ from the distribution F^0 and compute Δ_{n2}^*, the version of Δ_{n2} for the resampled data $\mathcal{X}^*$. To construct a test at level α, use Monte Carlo methods to compute the critical point $\hat{z}_\alpha$ defined by

$$P_{F^0}(\Delta_{n2}^* > \hat{z}_\alpha \mid \mathcal{X}) = \alpha,$$

and reject the null hypothesis that f is unimodal if $\Delta_{n2} > \hat{z}_\alpha$. Under mild regularity conditions, including that $\hat{c}$ is a consistent estimate of c under the null hypothesis, the test has asymptotically correct level (Cheng and Hall, 1998).

Cheng and Hall also show how their method can be applied to the more general problem of testing the null hypothesis that f has m modes against the alternative that f has $m+1$ modes, for $m \geq 2$. They conduct a simulation study that demonstrates that their test has very good level accuracy and greatly improved power. They compare their method with Silverman's (1981) critical bandwidth test and find that their method generally has better level accuracy under the null hypothesis and greater power under the alternative. It also avoids the problems of finding spurious modes in the tails of a distribution that the bandwidth test often encounters.

Cheng and Hall (1997) consider a second method of calibrating the test. Rather than explicitly estimating c they use a double bootstrap method, based on Silverman's (1981) resampling approach, that adjusts for c implicitly. This method also produces an asymptotically accurate test but it does not generalise to the problem

of testing the null hypothesis that f has two or more modes. Cheng and Hall's (1997) simulation study also suggests that the method that involves explicitly estimating c generally outperforms this double bootstrap technique. The method that we develop in Chapter 2 for calibrating Silverman's bandwidth test is closely related to this double bootstrap procedure for the excess mass test.

The most attractive feature of DIP and excess mass tests is that they are based directly on the empirical distribution function and do not require a density estimate to be calculated. A test that relies on a density estimate can be no more effective than the estimate on which it depends, since the test will inherit any undesirable properties that the estimator might have. We saw in Section 1.3 that the critical bandwidth test has problems with falsely detecting modes in the tails of a distribution and encounters difficulties with multimodal distributions whose modes are of widely varying shapes and sizes. These problems arise because the bandwidth test is based on a kernel density estimator with a global bandwidth, and a bandwidth that produces a good estimate in one region does not necessarily produce a good estimate in other regions. Another advantage to directly using the empirical distribution function to assess modality is that it does not require the specification of a bin width (e.g. Cox, 1966) or a smoothing parameter (e.g. Good and Gaskins, 1980).

However, the properties of the excess mass test that protect it from falsely identifying spurious modes in the tails can also make it insensitive to the presence of minor modes. The test is based on the amount of probability mass above a level λ. If the height of a minor mode is below the height of the trough between two major modes then λ will be chosen to be above the height of the minor mode and it will make no contribution to the excess mass statistic. Consequently the test will not be able to detect the presence of this smaller mode.

In spite of the difficulties associated with tests that rely on density estimation, the critical bandwidth test has proved to be more popular with practitioners than excess mass tests. This is partly due to the intuitive appeal of a method that depends on a simple density estimate, that can be plotted and examined, rather than a test that appeals to more mathematically abstract ideas such as Lebesgue measure.

1.6 The mode tree and mode existence test

Silverman's technique tests for the unimodality, bimodality or multimodality of a data set as a whole and is based on a kernel density estimate with a global bandwidth. Minnotte and Scott (1993) and Minnotte (1997) suggest a mode existence technique which tests the significance of each mode individually, using a different bandwidth for each potential mode. This approach can be advantageous since we are often interested in knowing whether the appearance of a concentration of data points in a sample represents a true mode of an underlying population, rather than knowing precisely how many modes there are in a population. Using different bandwidths for each potential mode allows a degree of local adaptivity which can be quite important, especially when modes occur on peaks of varying sizes.

The mode existence test relies on an exploratory, graphical tool developed by Minnotte and Scott (1993) which they called the *mode tree*. The mode tree is a graph which plots the mode locations for a kernel density estimate against the bandwidth at which the density estimate with those modes is calculated. It provides a simple yet informative format for displaying how modes split and new modes appear, as the bandwidth decreases and for which bandwidths these splits occur. Minnotte and Scott (1993) recommended the use of the Normal kernel because the monotonicity of the number of modes, as a function of the bandwidth, ensures that all modes found at a given level of h remain as h decreases.

Other features of density estimates besides the location of the modes can be added to the mode tree. These include the location of antimodes and points of inflection, and intervals which contain bumps could be shaded. An interval whose length is proportional to the size of the mode could also be shaded. Minnotte and Scott define the size of mode j, for a bandwidth h, as

$$M_j(h) = \int_{\hat{a}_j}^{\hat{b}_j} \left[\hat{f}_h(x) - \max\{\hat{f}_h(\hat{a}_j), \hat{f}_h(\hat{b}_j)\}\right]_+ dx\,,$$

where $\hat{a}_j$ and $\hat{b}_j$ are the locations of the left and right antimodes surrounding mode j. The quantity M_j is the probability mass of the mode above the higher of the two surrounding antimodes. In a sense M_j is the single mode equivalent to Müller and Sawitzki's (1991) excess mode functional, evaluated at the height of the the higher of the two surrounding antimodes. It differs from their statistic in being computed from a kernel density estimate rather than directly from the empirical distribution

function.

Minnotte and Scott (1993) and Minnotte (1997) propose using M_j as a test statistic to test the null hypothesis that mode j is an artefact of the sample against the alternative that mode j is a true feature of the population. Minnotte (1997) investigated the theoretical properties of M_j. He showed that M_j converges in probability to zero when the null hypothesis is true. When the alternative is true it converges to its true population value

$$\int_a^b \left[f(x) - \max\{f(a), f(b)\}\right]_+ dx\,,$$

where a and b are the locations of the left and right antimodes surrounding the mode. See the original paper for the exact rates of convergence and other technical details. These results show that the use of M_j as a test statistic produces a consistent test.

In implementing the test it is necessary to choose a bandwidth at which to compute $M_j(h)$. Minnotte (1997) shows that $M_j(h)$ is a non-increasing function of h and suggests choosing $\tilde{h}$ such that $M_j(\tilde{h})$ is as large as possible, in order to produce a test with maximum power. This is achieved by choosing $\tilde{h}$ as small as possible without the mode splitting into two new modes.

Following Silverman (1981), Minnotte and Scott (1993) suggest a resampling approach to assess the significance of M_j. They resample from $\hat{f}_{\tilde{h}}$ but with the mode that is being tested and at least one of the adjacent antimodes being replaced with a flat section. This approach can be viewed as being a single mode version of Hartigan and Hartigan (1985) and Müller and Sawitzki's (1991) approach to assessing the significance of their statistics based on comparison of properties with samples of Uniform random variables; see Section 1.5 for a discussion of these methods. These techniques are known to be even more conservative than Silverman's (1981) bootstrap technique. The p-values and power results reported in Minnotte's (1997) simulation study support this view.

The simulation study and Minnotte and Scott's (1993) application of the test to the stamp data (from Izenman and Sommer, 1988) show that when the modes occur on peaks of varying sizes the local, adaptive approach of their test has distinct advantages over the global approach of Silverman's test. In Section 2.4 and Chapter 4 we show that some of the problems associated with the non-adaptivity of using a global bandwidth can be alleviated by testing for the number of modes over a compact interval, which is a subset of the support of the density, rather than over

the whole real line.

1.7 Mode testing in higher dimensions

In the earlier sections of this chapter we have considered methods for assessing the modality of univariate populations. In this section we shall conduct a brief survey of methods that have been developed for investigating the modality of multivariate populations. Some of these methods are generalisations of univariate approaches, such as the work of Hartigan (1987) and Polonik (1995a) on multivariate excess mass estimates and Hartigan's (1988) adaption of his univariate DIP statistic into the multivariate SPAN statistic. Others, such as Hartigan and Mohanty's (1992) RUNT test and Rozál and Hartigan's (1994) MAP test, are novel approaches that can be applied to data of all dimensions but are designed particularly with higher dimensional data in mind.

Hartigan (1987) developed an excess mass approach for a density in two dimensions, independently of Müller and Sawitzki's (1987) work in one dimension. He examined the problem of estimating a convex density contour using excess mass to estimate the amount of probability mass above a given contour level. Hartigan used his method to develop a test for bimodality, with the test statistic being an estimate of the excess probability located in a secondary mode.

Polonik (1995a) examined estimating excess mass over a class of subsets of d-dimensional Euclidean space, and used this estimate to develop a generalisation of the excess mass statistic for multimodality (Müller and Sawitzki, 1991), for d-dimensional populations. Polonik (1995b) looked at the closely related problem of estimating a d-dimensional density under shape restrictions on the density contour clusters. By choosing these shape restrictions appropriately it is possible to model various features of the data, including multimodality.

Hartigan (1988) developed the SPAN test as a multidimensional generalisation of the DIP test. The SPAN test uses distribution functions defined on rooted minimal spanning trees instead of the usual multivariate distributions. The minimum spanning tree is a tree of minimum total length linking all data points. Given a root node r, the distance from root r defines a partial ordering on the nodes of the tree and F_{nr} an empirical distribution function corresponding to root r. The statistic SPAN(F_{nr}) is the minimum spanning tree version of the DIP, that is, it is

the maximum difference between the EDF and it closest unimodal approximant.

Given a sample $X_1, \ldots, X_n$, the empirical distribution P_n gives probability $1/n$ to each sample point. The SPAN for this distribution is defined to be

$$\text{SPAN}(P_n) = \min_r \text{SPAN}(F_{nr}),$$

that is, the minimum distance between the EDF and its closest unimodal approximant, minimised over all choices of the root node. The computation of the SPAN statistic is quite involved and the algorithm takes between n^2 and n^3 steps.

Hartigan conjectures that if the sample is from a continuous unimodal density f then $\text{SPAN}(P_n) \to 0$, as $n \to \infty$, and that when f is not unimodal, $\text{SPAN}(P_n)$ converges in distribution to a positive random variable. These results have been proven for the one-dimensional DIP statistic. The paper gives the 95% point of the SPAN distribution for up to 5 dimensions, based on samples drawn from a Uniform distribution on the sphere.

Hartigan and Mohanty (1992) propose the RUNT statistic as a more simply computed alternative to the SPAN statistic. It uses single linkage clustering for detecting the presence of multimodality in populations. Single linkage clusters on a set of points are the maximal connected sets in a graph constructed by linking all points within a given threshold distance of each other. The complete set of single linkage clusters is obtained from all graphs constructed using different threshold distances. As the threshold distance is decreased each cluster divides into two or more subclusters until each cluster consists of a single point. For each single linkage cluster the runt size is the number of points in its smallest subcluster. The RUNT statistic is the maximum runt size over all threshold distances. Large values of the RUNT suggest multimodality.

Justification of the test is based on the asymptotics of single linkage clusters; see Hartigan (1981). It there are at least two modes in the population density, then asymptotically, just one of the single linkage clusters will split into two clusters of points, one about each of the modes. The smaller of these two clusters will be the runt. If there is a single mode, then asymptotically, each cluster will divide into two clusters, the smaller of which will contain very few points. Therefore we would expect a larger number of points in the runt when the distribution is multimodal and thus a large value of the RUNT statistic should indicate the presence of multimodality.

Hartigan and Mohanty use samples from both the spherical Normal and Uniform distributions to calculate the percentage points of the RUNT statistic. Their

simulations show that the test has highest power for 3, 4 and 5 dimensions.

Rozál and Hartigan (1994) introduced a test based on minimal constrained spanning trees. They define a Minimal Ascending Path Spanning Tree (MAPST) which is the minimal spanning tree whose link lengths are non-increasing on the path to the root node, starting from any link. MAPSTs with more than one root are also defined to accommodate the possibility of multimodality. A multiple root MAPST is a minimal spanning tree constrained so that starting from any link there is a path to one of the root nodes satisfying the ascending path property. Rozál and Hartigan present algorithms for finding MAPSTs.

Nearest neighbour density estimates are used to develop a test statistic. Let $X_1, \ldots, X_n \in \mathcal{R}^d$ be a sample from a density f with an unknown number of modes. Let $r_k(x)$ be the Euclidean distance from x to the k-th nearest point. If c_d is the volume of the unit sphere in $\mathcal{R}^d$, then $V_k(x) = c_d r_k(x)$ is the d-dimensional volume of the d-dimensional sphere $S_k(x)$ of radius $r_k(x)$ centred at x. Of the n sample points, it is expected that about $nf(x)V_k(x)$ fall inside $S_k(x)$. Equating this expected number with the number k actually observed gives the nearest neighbour estimate of f at x,

$$\hat{f}(x) = \frac{k}{nV_k(x)} = \frac{k}{nc_d r_k(x)^d}. \tag{1.10}$$

The d-th root of the density estimate is inversely proportional to the distance of the k-th nearest point from x. Thus the density is highest where the points are closest together, indicating existence and location of modes. Taking logarithms of both sides of (1.10) gives

$$\log \hat{f}(x) = \log\left(\frac{k}{nc_d}\right) - d \log r_k(x), \tag{1.11}$$

indicating that $\log \hat{f}(x)$ is linear in the minus log distance to the k-th nearest point.

For each point in the sample $X_1, \ldots, X_n$ let $T_U(X_i)$ (U for unimodal) be a MAPST with a single root at X_i. Let $T_B(X_{i_1}, X_{i_2})$ (B for bimodal) be a MAPST with roots at X_{i_1} and X_{i_2}. Let $\rho_{X_i}(X_j)$ be the length of the link from X_j to the next mode $\nu_{X_i}(X_j)$ in the direction towards the root in $T_U(X_i)$. Define the functions $L_U(X_i)$ and $L_B(X_{i_1}, X_{i_2})$, which are the sum of the logarithms of the lengths of the links in $T_U(X_i)$ and $T_B(X_{i_1}, X_{i_2})$ respectively, by

$$L_U(X_i) = \sum_{j \neq i} \log \rho_{X_i}(X_j)$$

and

$$L_B(X_{i_1}, X_{i_2}) = \sum_{j \neq i_1, i_2} \log \left\{ \min_{i \in i_1, i_2} \rho_{X_i}(X_j) \right\} + \log D(X_{i_1}, X_{i_2}),$$

where $D(X_{i_1}, X_{i_2})$ is the distance between the subclusters belonging to the roots X_{i_1} and X_{i_2} in $T_B(X_{i_1}, X_{i_2})$.

Equation (1.11) suggests that $-\log \rho_{X_i}(X_j)$ is roughly proportional to the log density at X_j. Thus, when f is unimodal $-L_U(X_i)$ should be roughly proportional to the log-likelihood $\sum_{j \neq i} \log f(X_j)$. When f is bimodal some of the distances $\rho_{X_i}(X_j)$ will be substantially larger than the nearest neighbour distances. Then $L_U(X_i)$ should be larger than minus the log-likelihood. However, under bimodality $-L_B(X_{i_1}, X_{i_2})$ will be close to the log-likelihood.

The MAP statistic for bimodality is defined by

$$\text{MAP} = \min_i L_U(X_i) - \min_{i_1, i_2} L_B(X_{i_1}, X_{i_2}).$$

The quantities L_U and L_B are minimised when the roots are chosen to be close to the modes of f. Under unimodality, except for small links near the roots, the unimodal and bimodal trees will be nearly identical and the value of the MAP will be near zero. If f is bimodal the root of the minimal unimodal tree T_U will be near one of the modes and the links of T_U that connect points belonging to the other mode will be considerably longer than those links in the minimal bimodal tree T_B that connect those points into T_B. There will be a considerable difference between the minima of L_U and of L_B; the MAP will be bounded away from zero.

The MAP test can be extended to testing the null hypothesis that f has two or more modes. The significance of the test can be determined by comparing values from samples taken from a reference distribution or by using a resampling method proposed by Rozál and Hartigan (1994).

1.8 Finite mixture distributions

In the introduction to this chapter we argued that if the density connected with a population is multimodal then this suggests that the underlying population is not homogeneous and that the population can be considered to be made up of a number of more homogeneous components. Traditionally such populations have been modelled by a finite mixture of distributions, usually unimodal distributions.

Finite mixture densities have the form

$$f(x) = \sum_{j=1}^{k} \pi_j f_j(x|\theta_j)\,, \tag{1.12}$$

where $\pi_j > 0$ for $j = 1, \ldots, k$, $\sum_{j=1}^{k} \pi_j = 1$ and $f_1, \ldots, f_k$ are probability density functions. The parameters $\pi_1, \ldots, \pi_k$ are usually known as the *mixing proportions* and $f_1, \ldots, f_k$ are known as the *component densities.*

While the presence of multimodality is indicative of a mixture distribution not all mixtures are multimodal. For example, an equal mixture of two Normal distributions with a common variance will only be bimodal if the distance between the means of the two Normals is greater than twice the standard deviation of the component densities. When the components of a mixture are sufficiently well separated the mixture density will be multimodal. Heavily overlapping mixture components will tend to induce inflection points and skewness in the mixture densities, rather than multimodality.

Hence knowing the number of modes in a density will provide a lower bound on the number of components in a mixture. We shall see later in this section that determining the number of components in a mixture can be very awkward. In addition to this awkwardness, the use of mixtures generally requires the assumption of a parametric model for each of the components. The assessment of the number of components in a mixture can be heavily influenced by the validity of the particular parametric model. For example, if it is assumed that the components of a mixture are Normal but they are actually skewed then each component might require a mixture of two or more Normals to adequately describe it. This would result in the assessment of the number of components, based on the parametric model, to overestimate the number of homogeneous subpopulations contained in a population.

Finite mixture distributions have a wide variety of applications and have been extensively studied. Three monographs on the topic are Everitt and Hand (1981), Titterington, Smith and Makov (1985) and McLachlan and Basford (1988).

In applications the components densities in (1.12) usually take the same parametric form, but this is not a requirement. When the component densities are all Normals the model is said to be a *Normal mixture*. For the rest of this section we shall concentrate on Normal mixtures. For a Normal mixture (1.12) can be written

as

$$f(x) = \sum_{j=1}^{k} \pi_j \frac{1}{\sigma_j} \phi\left(\frac{x - \mu_j}{\sigma_j}\right) , \tag{1.13}$$

where ϕ is the standard Normal density function and μ_j and σ_j are the mean and standard deviation of the jth component, respectively.

The class of mixture distributions is an extremely rich one and if k and θ_i are unrestricted then any continuous density can be arbitrarily closely approximated by a Normal mixture. In fact, the Normal kernel density estimator (see Section 1.3) is a Normal mixture. This can be seen by setting $k = n$, $\pi_j = 1/n$, $\mu_j = X_j$ and $\sigma_j = h$ in equation (1.13). In practice there need to be some restrictions introduced in order for models of form (1.13) to be able to usefully estimate densities. For example, for kernel density estimates the bandwidth h is the only arbitrary parameter.

If we wish to estimate a density using a Normal mixture, based on a sample $X_1, \ldots, X_n$, there are two problems that need to be considered. Firstly, we need to determine k, the number of components in the mixture, and secondly, given k, we need to estimate the parameters π_j, μ_j and σ_j associated with each component. Neither of these two problems is as straightforward as it first appears. We shall consider the second problem first. The most common method for determining the parameters for a parametric model is maximum likelihood. For a Normal mixture the likelihood is

$$L(\pi, \mu, \sigma) = \prod_{i=1}^{n} \left[\sum_{j=1}^{k} \pi_j \frac{1}{\sigma_j} \phi\left(\frac{x_i - \mu_j}{\sigma_j}\right) \right] . \tag{1.14}$$

The maximum likelihood estimates π_j, $\hat{\mu}_j$ and $\hat{\sigma}_j$ of the parameters are those values that maximise (1.14). However the likelihood surface is littered with singularities. To see this, put $\mu_1 = X_1$, and note that $L \to \infty$ as $\sigma_1 \to 0$. In spite of these singularities, a strongly consistent maximum likelihood estimator can be obtained (Titterington *et al.*, 1985, Section 4.3.3) as long as we keep away from singularities on the boundary of the parameter space.

Since the maximiser of (1.14) has no explicit solution, it is necessary to use iterative, numerical methods to obtain maximum likelihood estimates. The EM algorithm (Dempster, Laird and Rubin, 1977) has become a popular method because of its ease of implementation, low storage requirements and robustness against poor initial values; see McLachlan and Basford (1988, Section 1.6). Hathaway (1985,

1986) developed a constrained version of the EM algorithm that prevents convergence to singularities. However in practice the likelihood surface can have many local maxima particularly when there is a large number of parameters to be fitted relative to the size of the data set. While Hathaway's algorithm reduces the chance of convergence to a local maximum, rather than the global one, the number of local maxima can make maximum likelihood estimation difficult. Other approaches used to avoid these problems include imposing the condition $\sigma_1 = \sigma_2 = \cdots = \sigma_k$ or using minimum distance estimation based on distribution functions (Titterington *et al.*, 1985, Section 4.5).

The other problem is determining k, the number of components in a mixture. This is a very difficult problem and no clear statistical procedure has been developed. The obvious generalised likelihood ratio test does not yield the usual asymptotic chi-squared distribution under the null. Consider testing the nested hypotheses

$$H_0 : f(x) = \frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)$$

and

$$H_1 : f(x) = \pi\frac{1}{\sigma_1}\phi\left(\frac{x-\mu_1}{\sigma_1}\right) + (1-\pi)\frac{1}{\sigma_2}\phi\left(\frac{x-\mu_2}{\sigma_2}\right).$$

The usual asymptotics do not apply since the regularity conditions on which they rely (see Cox and Hinkley, 1974, p. 281) are violated by the parameters of the null hypothesis being located on the boundary of the parameter space of the alternative hypothesis. In fact, rather than converging to a chi-squared distribution, Hartigan (1985) has shown that the likelihood ratio converges to infinity at a very slow rate. McLachlan (1987) has used the bootstrap to estimate the null distribution of the likelihood ratio test but this test is very computationally demanding and, in practice, lacks power because under the null the distribution of the test statistic has a long right tail (Roeder, 1994).

An alternative to formal testing procedures for determining the number of components in a mixture is the use of diagnostic graphical tools. These include the use of histograms and other density estimates, as well as quantile-quantile plots. These methods are described in Everitt and Hand (1981, Section 5.2.1). A more recent graphical approach was proposed by Roeder (1994). She showed that if the density of a mixture of two Normals is divided by a Normal density with the same mean and variance as the mixture, then the resulting function will always be bimodal. Roeder

developed a graphical tool and a formal testing procedure based on this result. Under the null hypothesis that $k = 1$, the proposed diagnostic can be approximated by a stationary Gaussian process while, under the alternative hypothesis, the components of the mixture will manifest themselves as modes in the diagnostic plot. The graphical tool involves examining a plot of the fluctuations in the process and comparing the amplitude of the fluctuations with asymptotic confidence bands derived under the null. The formal test borrows critical smoothing ideas from Silverman (1981) and is based on the amount of smoothing required to suppress the deviations from the stationary Gaussian process.

Roeder (1990) proposed a semiparametric density estimation technique that entails choosing a Normal mixture that maximises a function based on sample spacings. The estimation technique relies on the selection of a smoothing parameter. Roeder suggested using cross-validation to choose the smoothing parameter to obtain a point estimate of the density and provided a means of determining a confidence set of plausible densities. The confidence set consisted of the set of Normal mixtures fitted using a range of smoothing parameters. The boundaries of the smoothing parameter are determined by inverting a distribution-free goodness-of-fit statistic. By considering a range of values for the smoothing parameter and the accompanying confidence set of densities, ranging from smooth to rough, we can obtain a handle on the number of modes in the density.

Titterington *et al.* (1985, Section 4.4) outlined Bayesian methods that have been employed to fit mixture models. More recently, Escobar and West (1995) developed efficient simulation methods for approximating prior and posterior distributions for the number of components underlying an observed data set.

Chapter 2

Calibrating Silverman's test

2.1 Introduction

In this chapter we propose methods for overcoming some of the weaknesses of Silverman's (1981) critical bandwidth test for testing the null hypothesis that a distribution has j modes versus the alternative that it has $j+1$ or more modes. In Section 1.3 we identified two main difficulties with Silverman's test. The first of these problems was that the bootstrap part of the test does not consistently estimate the distribution of the critical bandwidth $\hat{h}_{\text{crit}}$, the test statistic. This results in a very conservative test. We propose a version of the test, in the important case $j = 1$ that produces a test with asymptotically correct level and which has greater power than the standard test. Bootstrap calibration techniques were first discussed by Hall (1986, 1987) and Loh (1987, 1991) in the context of improving the coverage accuracy of bootstrap confidence intervals.

The second problem was that the testing problem is often made more difficult by spurious modes arising from outlying data values. We suggest conducting the test over a compact interval rather than over the whole line. We study this modified version of the test and show that, asymptotically, this form of the test has correct level under quite general conditions.

The numerical results related to the work described in this chapter can be found in Chapter 3. In Chapter 4 we apply the various forms of the test to real data sets.

2.2 The bootstrap test

In Section 1.3 we described the critical bandwidth test as it was proposed by Silverman (1981) for testing the null hypothesis of j modes versus the alternative of $j+1$ or more modes. In this section we shall outline our version of the test for the important case of $j = 1$.

Given a data set $\mathcal{X} = \{X_1, \ldots, X_n\}$ from a distribution with unknown density f, we wish to test the null hypothesis H_0 that f has a single mode in the interior of a given closed interval $\mathcal{I}$, and no local minimum in $\mathcal{I}$, against the alternative hypothesis H_1 that f has more than one mode in $\mathcal{I}$. Following Silverman we construct the kernel density estimator

$$\hat{f}_h(x) = (nh)^{-1} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right),$$

and we take the kernel function K to be the standard Normal density function. We have seen in Section 1.3 that the number of modes of $\hat{f}_h$ on the whole line is always a nonincreasing function of the bandwidth h. Furthermore, $\hat{f}_h$ is unimodal for all sufficiently large h, and so

$$\hat{h}_{\text{crit}} = \inf\left\{h : \hat{f}_h \text{ has precisely one mode in } \mathcal{I}\right\} \tag{2.1}$$

is well-defined, at least when $\mathcal{I}$ is the whole line. The definition of $\hat{h}_{\text{crit}}$ for more general $\mathcal{I}$ will be discussed in Sections 2.4 and 2.6. See Section 1.3 for a discussion of the validity of using $\hat{h}_{\text{crit}}$ as a test statistic.

Let $\hat{f}_{\text{crit}}$ denote the version of $\hat{f}_h$ that we obtain by putting $h = \hat{h}_{\text{crit}}$. Conditional on $\mathcal{X}$, let $X_1^*, \ldots, X_n^*$ be a resample drawn from the distribution with density $\hat{f}_{\text{crit}}$, and put

$$\hat{f}_h^*(x) = (nh)^{-1} \sum_{i=1}^{n} K\left(\frac{x - X_i^*}{h}\right).$$

Let $\hat{h}_{\text{crit}}^*$ denote the bootstrap version of $\hat{h}_{\text{crit}}$, that is, the infimum of all bandwidths h such that $\hat{f}_h^*$ has precisely one mode. The test statistic is the bootstrap distribution of $\hat{h}_{\text{crit}}^*/\hat{h}_{\text{crit}}$, and an α-level test of H_0 against H_1 is to reject H_0 if, for an appropriate quantity λ_α,

$$P\left(\hat{h}_{\text{crit}}^*/\hat{h}_{\text{crit}} \le \lambda_\alpha \middle| \mathcal{X}\right) \ge 1 - \alpha.$$

This is a little different to the formulation of the test expounded in Section 1.3, but it is equivalent when $\lambda_\alpha \equiv 1$. The usual version of the test with $\lambda_\alpha \equiv 1$ is based

on the notion that the distribution of $U_n = P(\hat{h}^*_{\text{crit}} \leq \hat{h}_{\text{crit}}|\mathcal{X})$ is approximately Uniform (0,1), at least for large values of n. We show in Section 2.6 that this is not correct, even asymptotically, because the limiting distributions of $\hat{h}_{\text{crit}}$ and $\hat{h}^*_{\text{crit}}$ (conditional on $\mathcal{X}$) are different. This reformulation of the test allows it to be adjusted to correct for the non-uniformity of U_n to produce a test which is calibrated for level accuracy. In the next section we discuss methods for calibrating the test and in Chapter 3 we apply these techniques to numerically compute λ_α.

Another difference between this test and the one described in Section 1.3 is the resampling scheme that is used. Here the resampling is carried out from the distribution with density $\hat{f}_{\text{crit}}$, where as in Section 1.3 we used a modified version of this distribution which had been rescaled so that the first two moments of the resampled data were equal to the sample moments of the data. Essentially this is a finite-sample correction which is employed to reduce the conservatism of the test. Since in this chapter we are concerned with developing a test that has asymptotically correct level we can disregard this variance correction, which has no bearing on the first order asymptotics. In Chapter 3 we shall examine the performance of the calibrated test and Silverman's variance corrected test and we shall consider combining the two techniques to produce a test with good level accuracy for small to medium sized samples.

2.3 Consistency of the test

We shall show in Section 2.6 that, under H_0, the bootstrap distribution function,

$$\widehat{G}_n(\lambda) = P\big(\hat{h}^*_{\text{crit}}/\hat{h}_{\text{crit}} \leq \lambda|\mathcal{X}\big)\,,$$

converges weakly to a stochastic process $\widehat{G}$ whose distribution does not depend on unknowns. (Each realization of $\widehat{G}$ is a distribution function.) The finite-dimensional distributions of $\widehat{G}$ are absolutely continuous, and the realizations of $\widehat{G}$ are continuous functions with probability 1. Hence, for each α there exists a unique absolute constant λ_α such that $P\{\widehat{G}(\lambda_\alpha) \geq 1-\alpha\} = \alpha$. The value of λ_α is given in Section 3.2. Using these values, the bootstrap test is asymptotically correct, in that

$$P\Big\{P\big(\hat{h}^*_{\text{crit}}/\hat{h}_{\text{crit}} \leq \lambda_\alpha|\mathcal{X}\big) \geq 1-\alpha\Big\} \to P\big\{\widehat{G}(\lambda_\alpha) \geq 1-\alpha\big\} = \alpha\,. \qquad (2.2)$$

This approach to calculating λ_α is based entirely on asymptotic arguments. Therefore, though it produces an asymptotically correct test, the test might not

perform so well for finite samples. An alternative method is based on Monte Carlo methods and computes λ_α for a particular sample size n. This involves drawing samples of size n from a distribution whose density is unimodal, with its mode interior to $\mathcal{I}$, and applying the bootstrap test to each sample. This approach is potentially better able to correct for second-order effects, for example by taking into account the influence of a specific sample size.

However this method does have the drawback of requiring a choice of unimodal distribution from which to draw the sample. Since $\widehat{G}$ does not depend on any unknowns then asymptotically it does not matter which unimodal distribution we choose. For finite samples the value of λ_α will be affected by the choice of density but we expect that this effect will be small, on the basis of the asymptotic result. In Section 3.7 we calculate λ_α for a range of sample sizes by drawing the samples from a Normal distribution. We justify the choice of this distribution there.

We have seen that the test behaves as desired under H_0. In Section 2.6 we show that under H_1 the bootstrap distribution converges to a degenerate mass at the origin, in the sense that $P\{\widehat{G}_n(\lambda) \leq x\} \to 0$ for all $\lambda > 0$ and all $x < 1$. Therefore, asymptotically the null hypothesis will always be rejected when the alternative is true. Hence the bootstrap test is consistent.

2.4 Spurious modes and compact intervals

A practical problem that can arise when implementing the critical bandwidth test is the appearance of spurious modes in the tail of the kernel density estimate. These modes are simply artefacts of the sparseness of the data in the tails. In Section 1.3 we quoted a result of Mammen, Marron and Fisher (1992) which demonstrated that as $n \to \infty$, $\hat{h}_{\text{crit}}$ tends to zero. In particular, $\hat{h}_{\text{crit}}$ is of size $n^{-1/5}$ and its only dependence on the unknown density f is through the value of the density and its second derivative at the mode x_0. This result depends on f having bounded support (see Section 2.6 for the complete set of regularity conditions) and on $\mathcal{I}$ being the whole real line. If both the support of f and the interval $\mathcal{I}$ are unbounded then the properties of $\hat{h}_{\text{crit}}$ are generally determined by extreme values in the sample, not by the modes of f.

For example, if f is a unimodal density whose upper tail decreases like a constant multiple of $x^{-\beta-1}$ for some $\beta > 0$ (such as the density of a Student's t-distribution),

then the spacings between consecutive pairs of extremes in the sample $\mathcal{X}$ are of size $n^{1/\beta}$ (see for example, Galambos, 1978, Sections 2.1 and 2.8). Bandwidths of smaller size than $n^{1/\beta}$ will produce kernel density estimates with a sequence of modes in the upper tail. Therefore it may be proved that if $\mathcal{I}$ is unbounded on the right then $\hat{h}_{\text{crit}}$ diverges at least as fast as a constant multiple of $n^{1/\beta}$. For similar reasons, if f is a Normal density and $\mathcal{I}$ is unbounded on either the left or the right then $\hat{h}_{\text{crit}}$ cannot decrease to zero any faster than $(\log n)^{-1/2}$. Even if the support of f is compact there can be problems in practice. In cases where f decreases to zero sufficiently quickly at the extremities of its support, the size of $\hat{h}_{\text{crit}}$ is driven by the rate of decrease, since that determines the spacings of the extreme order statistics of f.

Izenman and Sommer (1988) suggest the use of a data transformation to reduce tail effects. Hall and Wood (1996) recommend solving this problem by truncating the density estimator away from the tails. In our framework this is roughly equivalent to taking $\mathcal{I}$ to be a compact interval within which f does not vanish.

We saw in Section 1.3 that when $\mathcal{I}$ is taken to be the real line, the number $N(h)$ of modes of $\hat{f}_h$ is a monotone decreasing function of h. This need not be true when $\mathcal{I}$ is a compact subset of the real line. The position of a mode can migrate slightly as h is altered; if it were just inside or just outside $\mathcal{I}$ for some h, it can switch to the other side as h is changed. This causes few practical problems, however, since the positions of modes are readily monitored as h is varied. Indeed, a turning point does not merge with other turning points as h is decreased, at least until a bandwidth is reached where the first three derivatives of $\hat{f}_h$ vanish simultaneously. This result is given as Theorem 2.1 in Section 2.6.

Moreover, provided f has no turning point on the boundary of $\mathcal{I}$, the probability that $N(h)$ is monotone in h, within a wide range of values of h, converges to 1 under relatively general conditions. This result and others enable us to establish the asymptotic validity of the bootstrap test under quite general conditions. See Section 2.6.

Izenman and Sommer (1988), Minnotte and Scott (1993), Hall and Wood (1996), Minnotte (1997) and Cheng and Hall (1998) all discuss the problem of spurious modes in the tails of a kernel density estimator that employs a global bandwidth. The work of Minnotte and Scott (1993) and Minnotte (1997) identifies another problem with mode testing using a global bandwidth. This is the problem of detecting modes located on peaks of varying sizes; see Section 1.6. They suggest overcoming

the problem by using a different bandwidth for each potential mode. Another approach is to choose $\mathcal{I}$ to cover different closed subsets of the support of f. This will allow the use of different bandwidths for different peaks in the density. Even though our bootstrap test is only valid for testing the null hypothesis that the density has one mode, by appropriately subdividing the support of f into smaller intervals we can make deductions about the total number of modes of f. We make use of this idea in the application of our technique to real data sets in Chapter 4.

2.5 Testing for j modes

The problem of testing the null hypothesis H_{0j} that f has precisely j modes in $\mathcal{I}$, against the alternative that it has $j+1$ or more modes there, is in principle similar for all values of j; see Silverman (1981). Indeed, defining

$$\hat{h}_{\mathrm{crit},j} = \inf\left\{h : \hat{f}_h \text{ has precisely } j \text{ modes in } \mathcal{I}\right\},$$

$n^{1/5}\,\hat{h}_{\mathrm{crit},j}$ converges in distribution under H_{0j} to $\max_{0\le i\le 2j-1}(c_i R_i)$, where $c_i = f(t_i)^{1/5}/|f''(t_i)|^{2/5}$, $t_1,\ldots,t_{2j-1}$ are the turning points of f in $\mathcal{I}$ (assumed to all satisfy $f''(t_i)\neq 0$), and $R_1,\ldots,R_{2j-1}$ are independent and identically distributed random variables with a distribution that we shall define at (2.5). The bootstrap distribution function,

$$\widehat{G}_n(\lambda|j) = P\big(\hat{h}^*_{\mathrm{crit},j}/\hat{h}_{\mathrm{crit},j} \le \lambda \big| \mathcal{X}\big),$$

again converges weakly to a stochastic process, $\widehat{G}(\cdot|j)$, but unless $j=1$ the distribution of the latter process depends on the $2j-2$ unknowns $\{c_i/c_1 : 2\le i\le 2j-1\}$. Therefore, if $j\ge 2$ then the bootstrap test cannot be calibrated by simply forming the ratio $\hat{h}^*_{\mathrm{crit},j}/\hat{h}_{\mathrm{crit},j}$.

One possible way of calibrating the test in the case $j\ge 2$ is to follow the ideas of Cheng and Hall (1998). The values of c_i can be estimated by $\hat{c}_i$, which could be obtained using kernel estimates of f and f''. Since the limiting distribution of $\hat{h}_{\mathrm{crit},j}$ only depends on the unknown density f through the c_i's, by resampling from a mixture of distributions whose values of c_i are equal to $\hat{c}_i$ we shall obtain a bootstrap test which is asymptotically accurate. See Cheng and Hall (1998) for the exact procedure and proof of its validity. We do not pursue this idea in this thesis.

2.6 Theoretical results

In this section we formally state the theoretical results used in the earlier sections of this chapter to describe properties of the bootstrap test. These results will be proven in Section 2.7. The work in this section is joint work with Professor P.G. Hall.

2.6.1 Separation of the modes of $\hat{f}_h$

In Section 1.3 we described the result of Silverman (1981) that demonstrated that when the Normal kernel is used the number of modes of a kernel density estimator is a nonincreasing function of the bandwidth. This fact greatly simplifies the calculation of the critical bandwidths $\hat{h}_{\text{crit}}$ and $\hat{h}^*_{\text{crit}}$, and adds to the appeal of using critical bandwidths as test statistics. However, Silverman's result only applies to the case where the number of modes on the whole line is addressed.

The following result shows that the Normal kernel has another desirable property that is useful when we address the problem of testing for the number of modes over a compact interval, rather than over the whole line. It shows that as the bandwidth decreases a turning point of $\hat{f}_h$ remains isolated – that is, it does not merge with other turning points – as least until a bandwidth is reached where the first three derivatives of $\hat{f}_h$ vanish simultaneously. This allows the position of the modes to be readily monitored. The ability to track modes as the bandwidth decreases enables the mode testing problem to be addressed within a compact interval, even if the density has unbounded support. See Section 2.4 for discussion of the benefits of this approach.

Theorem 2.1 *Assume that K is the standard Normal density. Given $h_1 > 0$, let x_{h_1} denote a point such that $\hat{f}'_{h_1}(x_{h_1}) = 0$ and $\hat{f}''_{h_1}(x_{h_1}) \neq 0$. As h is decreased through positive values less than h_1, x_h varies continuously through points x satisfying $\hat{f}'_h(x) = 0$, $\hat{f}''_h(x) \neq 0$ and $\text{sgn}\{\hat{f}''_h(x)\} = \text{sgn}\{\hat{f}''_{h_1}(x_{h_1})\}$, at least until a value $h_2 < h_1$ is encountered with the property that $\hat{f}'_{h_2}(x_{h_2}) = \hat{f}''_{h_2}(x_{h_2}) = \hat{f}'''_{h_2}(x_{h_2}) = 0$.*

2.6.2 Distribution of $\hat{h}_{\text{crit}}$ under H_0

We define a *level point* of f to be a real number t such that $f'(t) = 0$, and a *turning point* to be a level point such that $\text{sgn}\{f'(t+)\} = -\text{sgn}\{f'(t-)\}$. Under H_0, f has

a unique turning point t_0 in $\mathcal{I}$. Put

$$c = f(t_0)^{1/5} / |f''(t_0)|^{2/5}, \tag{2.3}$$

and let

$$Z(r,s) = r^{-3} \int K''(s+u)\, W(ru)\, du\,, \tag{2.4}$$

where K is the standard Normal kernel, W is a standard Wiener process, $r > 0$ and $-\infty < s < \infty$. Define

$$\begin{aligned} R \;=\; & \inf\{r > 0 : \text{the function } Z(r,s) + s \text{ changes sign} \\ & \qquad \text{exactly once in the range } -\infty < s < \infty\}\,, \end{aligned} \tag{2.5}$$

and let S be the unique point at which $Y(s) = Z(R,s) + s$ changes sign. (There exists another point S_1 such that $Y(S_1) = 0$, but $\operatorname{sgn}\{Y(S_1+)\} = \operatorname{sgn}\{Y(S_1-)\}$). The variation-diminishing property of the integral operator with kernel K'' ensures that with probability 1 the number of sign changes of $Z(r,s)+s$ is a right-continuous, nonincreasing function of r; see Schoenberg (1950) and Silverman (1981). Similarly, if $\mathcal{I} = (-\infty, \infty)$ then the property ensures that $\hat{h}_{\text{crit}}$ is well-defined by (2.1).

The next result describes the limiting distribution of $\hat{h}_{\text{crit}}$. It reveals that asymptotically, the value of $\hat{h}_{\text{crit}}$ only depends on the unknown density f through c. Following Mammen, Marron and Fisher (1992) we impose regularity conditions that require f to be compactly supported, and allow $\mathcal{I}$ to be the whole real line. Theorem 2.5 will address alternative settings, where $\mathcal{I}$ is compact and f may be infinitely supported.

Theorem 2.2 *Assume that f is supported on a compact interval $\mathcal{S} = [a,b]$, and has two continuous derivatives there; that it has a mode t_0, giving a local maximum, in the interior of $\mathcal{S}$, with $f''(t_0)\, f(t_0) \neq 0$; that f has no other level points in $\mathcal{S}$; and that $f'(a+) > 0$ and $f'(b-) < 0$. Take $\mathcal{I} = (-\infty, \infty)$. Then, $n^{1/5}\, \hat{h}_{\text{crit}}$ converges in distribution to cR as $n \to \infty$.*

2.6.3 Bootstrap distribution of $\hat{h}^*_{\text{crit}}/\hat{h}_{\text{crit}}$ under H_0

For each n it is possible to construct the Wiener process W above such that, with R defined at (2.5), we have $n^{1/5}\hat{h}_{\text{crit}} = cR + o_p(1)$. Let W^* denote a second standard

Wiener process, independent of W, and let Z^* have the definition at (2.4) except that W there should now be replaced by W^*. Put

$$\begin{aligned}\zeta^*(r,s) &= Z^*(r,s) + r^{-1}\,R\,Z\big(R, S + R^{-1}rs\big) + r^{-1}RS + s\,,\\ R^* &= \inf\{r > 0 : \text{the function } \zeta^*(r,s) \text{ changes sign}\\ &\qquad\qquad \text{exactly once in the range } -\infty < s < \infty\}\,.\end{aligned}$$

The next theorem describes the limiting bootstrap distribution of $\hat h^*_{\rm crit}$. It shows that, like $\hat h_{\rm crit}$, $\hat h^*_{\rm crit}$ is of order $n^{-1/5}$. In addition it demonstrates that, like the limiting distribution of $\hat h_{\rm crit}$, the limiting bootstrap distribution of $\hat h^*_{\rm crit}$ only depends on f through c. Since the asymptotic distributions of $\hat h_{\rm crit}$ and $\hat h^*_{\rm crit}$ depend on unknowns only through the same constant c, then by forming the ratio $\hat h^*_{\rm crit}/\hat h_{\rm crit}$ we obtain a quantity whose limiting distribution is independent of the unknown density f.

Theorem 2.3 *Assuming the conditions of Theorem 2.2,*

$$\sup_{-\infty<x<\infty} \big|P\big(n^{1/5}\,\hat h^*_{\rm crit} \le cx \big| \mathcal{X}\big) - P\big(R^* \le x \big| W\big)\big| \to 0$$

in probability as $n \to \infty$.

Taking $x = n^{1/5}\,c^{-1}\,\lambda\,\hat h_{\rm crit} = \lambda R + o_p(1)$ in Theorem 2.3, we deduce that the stochastic process $\widehat G_n$ defined by $\widehat G_n(\lambda) = P(\hat h^*_{\rm crit}/\hat h_{\rm crit} \le \lambda | \mathcal{X})$ converges weakly to $\widehat G$, where $\widehat G(\lambda) = P(R^*/R \le \lambda | W)$. By definition of the distributions of R and R^*, the distribution of the stochastic process $\widehat G$ does not depend on unknowns. This confirms the asymptotic accuracy of the calibrated test proposed in Section 2.3, in particular the validity of formula (2.2) under the null hypothesis.

2.6.4 Distribution of $\hat h_{\rm crit}$ and $\hat h^*_{\rm crit}/\hat h_{\rm crit}$ under H_1

We show that $\hat h_{\rm crit}$ tends to be much larger, and $\hat h^*_{\rm crit}/\hat h_{\rm crit}$ much smaller, under H_1 than under H_0. This means that the test is consistent, in the sense that asymptotically the null hypothesis will always be rejected when the alternative is true.

Theorem 2.4 *Assume the conditions of Theorem 2.2, except that where before f had just one local maximum in the interior of $\mathcal{S}$, we now ask that it have m turning points $t_1, \ldots, t_m$ there, that $2 \le m < \infty$ and each $f(t_i)\,f''(t_i) \ne 0$. Then (a) there exists a constant $c > 0$ such that $P(\hat h_{\rm crit} > c) \to 1$, and (b) for each $\lambda > 0$, $\widehat G_n(\lambda) \to 1$ in probability.*

2.6.5 Alternative regularity conditions

Here we let $\mathcal{I}$ be a proper subset of the support of f, and show that in such cases the earlier results continue to be valid in an asymptotic sense. The main difference from taking $\mathcal{I}$ to be the whole real line is that we no longer require f to have bounded support.

By way of notation, let ϵ_n be a sequence of positive constants converging to zero, let $\delta > 0$ be fixed, put $\mathcal{H}_n = [\epsilon_n n^{-1/5}, \delta]$, let $\mathcal{H}'_n$ denote the set of all $h \in \mathcal{H}_n$ such that $\hat{f}_h$ has precisely one mode in $\mathcal{I}$, and redefine $\hat{h}_{\text{crit}} = \inf \mathcal{H}'_n$ if $\mathcal{H}'_n$ is nonempty, and $\hat{h}_{\text{crit}} = \delta$ otherwise. Let $N(h)$ denote the number of modes of $\hat{f}_h$ within $\mathcal{I}$.

Theorem 2.5 *Assume that $\mathcal{I}$ is compact, that f'' is bounded and continuous in an open interval containing $\mathcal{I}$, that f has only a finite number of level points t in $\mathcal{I}$, at each of which $f(t)\, f''(t) \neq 0$, and that neither endpoint of $\mathcal{I}$ is a level point of f. Then, if $\epsilon_n \to 0$ sufficiently slowly and $\delta > 0$ is sufficiently small, (a) the probability that $N(\cdot)$ is nonincreasing on $\mathcal{H}_n$ converges to 1, (b) under H_0, the probability that $\mathcal{H}'_n$ is not empty converges to 1, and (c) the conclusions of Theorems 2.2–2.4 apply for the above definition of $\hat{h}_{\text{crit}}$. (The analogues of Theorems 2.2 and 2.3 require that we assume additionally that f have one mode interior to $\mathcal{I}$, and no other turning points there; and the analogue of Theorem 2.4 requires f to have at least two modes interior to $\mathcal{I}$.)*

It follows that if we restrict attention to bandwidths that are not too small then, with probability tending to 1 as n increases, the calibration methods suggested in Section 2.2 apply in the case of testing for the number of modes on a compact interval.

2.7 Technical arguments

In this section we prove the theorems that were stated in the previous section. The work in this section is joint work with Professor P.G. Hall.

2.7.1 Proof of Theorem 2.1

By using properties of derivatives of the Normal density it can be shown that $(\partial/\partial h)\, \hat{f}_h^{(i)}(x) = h\, \hat{f}_h^{(i+2)}(x)$. Let x_h denote a turning point of $\hat{f}_h$, so that $\hat{f}'_h(x_h) \equiv 0$;

and put $x'_h = (\partial/\partial h)\, x_h$. In this notation,

$$0 = \frac{\partial}{\partial h}\, \hat{f}'_h(x_h) = \frac{\partial}{\partial h}\, \hat{f}'_h(x)\Big|_{x=x_h} + \hat{f}''_h(x_h)\, x'_h = h\, \hat{f}^{(3)}_h(x_h) + \hat{f}''_h(x_h)\, x'_h\,.$$

Therefore, $x'_h = -h\, \hat{f}^{(3)}_h(x_h)/\hat{f}''_h(x_h)$, whence

$$\frac{\partial}{\partial h}\, \hat{f}^{(i)}_h(x_h) = h\, \hat{f}^{(i+2)}_h(x_h) - h\, \hat{f}^{(i+1)}_h(x_h)\, \hat{f}^{(3)}_h(x_h)\, \{\hat{f}''_h(x_h)\}^{-1}\,. \tag{2.6}$$

As we alter h, x_h varies continuously through values of x such that $\hat{f}'_h(x) = 0$. Suppose that $\hat{f}''_{h_1}(x_{h_1}) < 0$ for some h_1. If it is not true that $\hat{f}''_h(x_h) < 0$ for all $0 < h < h_1$, then the quantity $h_2 = \sup\{h < h_1 : \hat{f}''_h(x_h) = 0\}$ is well-defined, and $0 < h_2 < h_1$. Suppose $\hat{f}^{(3)}_{h_2}(x_{h_2}) \neq 0$. Taking $i = 2$ in (2.6), we have that $(\partial/\partial h)\, \hat{f}''_h(x_h) \to +\infty$ as $h \downarrow h_2$. This means that $\hat{f}''_h(x_h)$ decreases through negative values as $h \downarrow h_2$, contradicting the hypothesis that $\hat{f}''_{h_2}(x_{h_2}) = 0$.

Similarly, if $\hat{f}''_{h_1}(x_{h_1}) > 0$ for some h_1 then $\hat{f}'_h(x_h) > 0$ for all $h < h_1$, unless or until a point is reached at which the first three derivatives of $\hat{f}_h$ vanish simultaneously; call this event $\mathcal{E}$. The result proved in the previous paragraph means that we may track a mode as h decreases, and it keeps the mode property. It will be uniquely defined unless it merges with another mode. In this event, since there must be at least one local minimum between the two modes, the local minimum must merge with the two modes at the same time as the modes merge, which is precluded by the result in the first sentence of this paragraph unless or until h has become so small that $\mathcal{E}$ occurs. Hence, each mode (and similarly, each local minimum) remains distinct as we decrease h, unless or until $\mathcal{E}$ occurs.

2.7.2 Proof of Theorem 2.2

The main idea in the proof of this theorem is to approximate $\hat{f}'_h(t)$ by a Gaussian process. It is based on the ideas of Mammen, Marron and Fisher (1992) . Mammen *et al.* have shown that under the conditions of Theorem 2.2 $\hat{h}_{\text{crit}}$ is of order $n^{-1/5}$ (Corollary 2.1) and that for bandwidths of this order a kernel density estimator will only have turning points within a neighbourhood of size $n^{-1/5}$ of a turning point of the underlying density that is being estimated (Theorem 3.1). If we put $h_0 = n^{-1/5}$

then these results can be presented as

$$\lim_{C_1 \downarrow 0, C_2 \uparrow \infty} \liminf_{n \to \infty} P\big(C_1 h_0 \leq \hat{h}_{\text{crit}} \leq C_2 h_0\big) = 1\,, \tag{2.7}$$

and

$$\lim_{C_1 \downarrow 0, C_2 \uparrow \infty} \limsup_{n \to \infty} P\big\{\hat{f}_h'(x) = 0 \text{ for some } h \in \big[C_1 h_0,\, C_2 h_0\big] \text{ and } x \notin \big[t_0 - C_2 h_0,\, t_0 + C_2 h_0\big]\big\} = 0\,. \tag{2.8}$$

The proof of these results relies on the embedding of Kómlos, Major and Tusnády (1975) that approximates the empirical distribution function, $\widehat{F}$, by $F + n^{-1/2} W^0$, where W^0 is a standard Brownian bridge and F is the distribution function associated with f. This approximation has been used by Mammen *et al.* and Silverman (1978) to approximate

$$X(h,t) = \hat{f}_h'(t) - E\{\hat{f}_h'(t)\}$$

by

$$Z_1(h,t) = -n^{-1/2} h^{-3} \int K''\left(\frac{t-x}{h}\right) W^0\{F(x)\}\, dx. \tag{2.9}$$

The following holds with probability $1 - o(n^{-1})$:

$$\Big\{ \sup_t \big|X(h,t) - Z_1(h,t)\big| < c_1 (\log n) n^{-2/5} \Big\}, \tag{2.10}$$

for c_1 large enough.

During the proof of Theorem 4 of Mammen *et al.* it is shown that if $h^\dagger$ denotes the infimum of values of h such that

$$Z_1(h,t) + f''(t_0)(t - t_0) \tag{2.11}$$

(as a function of t) has precisely one zero, then $\hat{h}_{\text{crit}} = h^\dagger + o_p(h_0)$ as $n \to \infty$.

We may replace $W^0\{F(x)\}$ by $D(x) = W^0\{F(x)\} - W^0\{F(t_0)\}$ in the integrand in equation (2.9) without affecting the value of the integral. Using properties of the modulus of continuity of a Gaussian process (see e.g. Garsia, 1970; Silverman, 1978) and standard properties of a Brownian bridge we may prove that

$$D(t_0 + hs) = W^0\{F(t_0) + hs\, f(t_0)\} - W^0\{F(t_0)\} + O_p\big\{h_0 (\log n)^{1/2}\big\}$$

uniformly in $|hs| = O(h_0)$. Therefore, defining a standard Weiner process,

$$W_1(u) = \{h_0 f(t_0)\}^{-1/2} \big[W^0\{F(t_0) - h_0 u\, f(t_0)\} - W^0\{F(t_0)\}\big],$$

and

$$Z_2(\rho, s) = f(t_0)^{1/2}\, \rho^{-3} \int K''(s+u)\, W_1(\rho u)\, du\,,$$

we have

$$Z_1(h, t_0 + hs) = h\, Z_2(h/h_0, s) + O_p\Big\{ \big(h_0^3 \log n\big)^{1/2} \Big\}\,, \tag{2.12}$$

uniformly in $h \in [C_1 h_0,\, C_2 h_0]$ and $s \in [-C_2 h_0/h,\, C_2 h_0/h]$ for any $0 < C_1 < C_2 < \infty$. Noting (2.7) and (2.8) we see that this limitation on h and s is no impediment.

So far we have shown that (2.11) is asymptotically equivalent to

$$h\, Z_2(h/h_0, s) + f''(t_0)(t - t_0), \tag{2.13}$$

but we also need to show that (2.11) and (2.13) have the same number of zeros. This can be achieved by proving that (2.13) is a monotone function of t in the neighbourhood of t_0. To do this we apply the following result (Lemma 4) of Mammen *et al.*

For every $a_n = o(1)$ and c_0:

$$\liminf_{n\to\infty} P\big\{ \big|Y'(t)\big| > a_n \text{ for } t \in U_n \text{ such that } \big|Y(t)\big| \le h_0 a_n \big\} = 1, \tag{2.14}$$

where

$$Y(t) = E\{\hat{f}_h'(t)\} + Z_1(h, t)$$

and

$$U_n = \{t : |t - t_0| \le c_0\, h_0\, (\log n)^{1/2}\}.$$

Mammen *et al.* have shown that the quantity at (2.11) has the same number of zeros as $\hat{f}_h'(t)$ so it just remains to show that replacing $Z_1(h, t)$ by $h\, Z_2(h/h_0, s)$ does not change the number of zeros. This can be done by applying (2.14), with $a_n = c\, h_0^{1/2}\, (\log n)^{1/2}$, and observing from (2.13) that

$$\big|h\, Z_2(h/h_0, s) - Z_1(h, t_0 + hs)\big| \le h_0 a_n.$$

Following our derivation of (2.12) (and noting that $ds/dt = h^{-1}$) we may similarly show that

$$h^{-1}\, (\partial/\partial s)\, Z_1(h, t_0 + hs) = (\partial/\partial s)\, Z_2(h/h_0, s) + O_p\big\{ h_0^{1/2}\, (\log n)^{1/2} \big\},$$

and hence that

$$\big|(\partial/\partial s)\, Z_2(h/h_0, s) - h^{-1}\, (\partial/\partial s)\, Z_1(h, t_0 + hs)\big| \le a_n.$$

Therefore, (2.13) not only tends to zero in the neighbourhood of t_0 but its derivative (with respect to t) is also bounded away from zero, which implies that (2.13) is a monotone function of t. This means that (2.13) has one, and only one, zero crossing in the neighbourhood of t_0. Hence, $h^\dagger/h_0 = R^\dagger + o_p(1)$, where $R^\dagger$ is the infimum of values of ρ such that $Z_2(\rho, s) + f''(t_0)s$ (as a function of s) has precisely one zero.

Put $c = \{f(t_0)/|f''(t_0)|^2\}^{1/5}$ and $W(t) = \operatorname{sgn}\{f''(t_0)\}\, c^{-1/2} W_1(ct)$, and for this W, define (as in Section 2.6.2)

$$Z(r, s) = r^{-3} \int K''(s + u)\, W(ru)\, du.$$

Then W is a standard Weiner process, and $Z_2(\rho, s)/f''(t_0) = Z(\rho/c, s)$. Therefore, for

$$R = \inf\{r > 0 : \text{the function } Z(r, s) + s \text{ changes sign exactly once in the range } -\infty < s < \infty\},$$

$R^\dagger = cR$, and so $\hat{h}_{\text{crit}} = h^\dagger + o_p(h_0) = h_0 R^\dagger + o_p(h_0) = h_0 cR + o_p(h_0)$, which completes the proof of Theorem 2.2.

2.7.3 Proof of Theorem 2.3

Since Theorem 2.3 is the bootstrap version of Theorem 2.2, the proof of Theorem 2.3 proceeds similarly to the previous proof. We begin by stating the bootstrap versions of equations (2.7)-(2.9). They are,

$$\lim_{C_1 \downarrow 0, C_2 \uparrow \infty} \liminf_{n \to \infty} P\big(C_1 h_0 \le \hat{h}^*_{\text{crit}} \le C_2 h_0 \,\big|\, \mathcal{X}\big) = 1\,, \tag{2.15}$$

$$\lim_{C_1 \downarrow 0, C_2 \uparrow \infty} \limsup_{n \to \infty} P\big\{\hat{f}^{*'}_h(x) = 0 \text{ for some } h \in [C_1 h_0, C_2 h_0] \text{ and } x \notin [\hat{t}_0 - C_2 h_0, \hat{t}_0 + C_2 h_0] \,\big|\, \mathcal{X}\big\} = 0\,, \tag{2.16}$$

and

$$Z^*_1(h, t) = -n^{-1/2} h^{-3} \int K''\left(\frac{t - x}{h}\right) W^{*0}\{F^*(x)\}\, dx, \tag{2.17}$$

where F^* is the distribution function associated with the density $\hat{f}_{\text{crit}}$, and W^{*0} is (conditional on $\mathcal{X}$) a standard Brownian bridge.

Similarly to the previous proof $Z_1^*(h,t)$ is an approximation to $X^*(h,t) = \hat{f}_h^{*'}(t) - E\{\hat{f}_h^{*'}(t) \mid \mathcal{X}\}$, and the following holds, with probability, conditional on $\mathcal{X}$, tending to 1:

$$\Big\{ \sup_t \big| X^*(h,t) - Z_1^*(h,t) \big| < c_1 (\log n) n^{-2/5} \Big\}, \tag{2.18}$$

for c_1 large enough.

Continuing through the argument of the previous proof we have,

$$W_1^*(u) = \{h_0 \hat{f}_{\text{crit}}(\hat{t}_0)\}^{-1/2} \left[W^{*0}\{F^*(t_0) - h_0 u\, \hat{f}_{\text{crit}}(\hat{t}_0)\} - W^{*0}\{F^*(\hat{t}_0)\} \right],$$

which is a standard Weiner process, conditional on $\mathcal{X}$, and

$$Z_2^*(\rho, s) = -f(t_0)^{1/2}\, \rho^{-3} \int K''(s+u)\, W_1^*(\rho u)\, du\,. \tag{2.19}$$

Notice that (2.19) contains the term $f(t_0)^{1/2}$ rather than $\hat{f}_{\text{crit}}(\hat{t}_0)^{1/2}$. This can be done since

$$\begin{aligned} \left| \hat{f}_{\text{crit}}(\hat{t}_0) \right|^{1/2} &= \left| f(t_0) \right|^{1/2} \left| 1 + \{\hat{f}_{\text{crit}}(\hat{t}_0) - f(t_0)\}/f(t_0) \right|^{1/2} \\ &= \left| f(t_0) \right|^{1/2} \left| 1 + O_p(h_0) \right|^{1/2}, \end{aligned}$$

and hence this substitution will only induce an error term of size $O_p(h_0^{3/2})$ in the following equation. This term is negligible compared to the other error term, $O_p\Big\{ \left(h_0^3 \log n\right)^{1/2} \Big\}$, so the analogue of (2.12) is

$$Z_1^*(h, \hat{t}_0 + hs) = h\, Z_2^*(h/h_0, s) + O_p\Big\{ \left(h_0^3 \log n\right)^{1/2} \Big\}, \tag{2.20}$$

uniformly in $h \in [C_1 h_0,\, C_2 h_0]$ and $s \in [-C_2 h_0/h,\, C_2 h_0/h]$ for any $0 < C_1 < C_2 < \infty$.

From this point the argument of the proof starts to differ from that of the previous one. In that proof we used $h\, Z_2(h/h_0, s) + f''(t_0)(t - t_0)$ as an approximation to $\hat{f}_h'(t)$. However, in the bootstrap version we need to avoid expressions involving $\hat{f}_{\text{crit}}''(\hat{t}_0)$, since it is not a consistent estimate of $f''(t_0)$. Bandwidths of a size larger that $n^{-1/5}$ are required for consistently estimating second derivatives.

Instead we define

$$Y^*(t) = E\{\hat{f}_h^{*'}(t) \mid \mathcal{X}\} + Z_1^*(h,t)\,, \tag{2.21}$$

and notice from (2.18) that

$$Y^*(t) = \hat{f}_h^{*\prime}(t) + O_p\big(h_0^2 \log n\big) \tag{2.22}$$

and

$$E\{\hat{f}_h^{*\prime}(t) \mid \mathcal{X}\} = \hat{f}'_{\rm crit}(t) + O_p(h_0^2)\,. \tag{2.23}$$

From (2.10) we have

$$\begin{aligned}\hat{f}'_{\rm crit}(\hat{t}_0 + hs) = Z_1\Big[\hat{h}_{\rm crit}, t_0 + \hat{h}_{\rm crit}\big\{(h/\hat{h}_{\rm crit}) + (s + \hat{s}_h)\big\}\Big] \\ + E\Big[\hat{f}'_{\rm crit}\big\{t_0 + h(s + \hat{s}_h)\big\} \mid \mathcal{X}\Big] + O_p\big(h_0^2 \log n\big),\end{aligned} \tag{2.24}$$

where $\hat{s}_h = (\hat{t}_0 - t_0)/h$. Rewriting (2.12) yields

$$\begin{aligned}Z_1\Big\{\hat{h}_{\rm crit}, t_0 + \hat{h}_{\rm crit}(h/\hat{h}_{\rm crit})(s + \hat{s}_h)\Big\} = \\ \hat{h}_{\rm crit} Z_2\Big\{\hat{h}_{\rm crit}/h_0, (h/\hat{h}_{\rm crit})(s + \hat{s}_h)\Big\} + O_p\Big\{\big(h_0^3 \log n\big)^{1/2}\Big\}\,.\end{aligned} \tag{2.25}$$

By taking a Taylor series expansion about t_0 and using standard properties of the normal kernel, K, we may show that

$$E\Big[\hat{f}'_{\rm crit}\big\{t_0 + h(s + \hat{s}_h)\big\} \mid \mathcal{X}\Big] = h(\hat{s}_h + s) f''(t_0) + O_p(h_0^2)\,. \tag{2.26}$$

Substituting (2.24), (2.25) and (2.26) into (2.23) yields

$$\begin{aligned}E\{\hat{f}_h^{*\prime}(t) \mid \mathcal{X}\} = \hat{h}_{\rm crit} Z_2\Big\{\hat{h}_{\rm crit}/h_0, (h/\hat{h}_{\rm crit})(s + \hat{s}_h)\Big\} \\ + h(\hat{s}_h + s) f''(t_0) + O_p(h_0^2 \log n)\,.\end{aligned} \tag{2.27}$$

Therefore, from equations (2.20), (2.22) and (2.27) we may obtain

$$\begin{aligned}\hat{f}_h^{*\prime}(t) = h\, Z_2^*(h/h_0, s) + \hat{h}_{\rm crit} Z_2\Big\{\hat{h}_{\rm crit}/h_0, (h/\hat{h}_{\rm crit})(s + \hat{s}_h)\Big\} \\ + h(\hat{s}_h + s) f''(t_0) + O_p\Big\{\big(h_0^3 \log n\big)^{1/2}\Big\}\,.\end{aligned} \tag{2.28}$$

If we define $Z^*(r, s)$ as in Section 2.6.3 and note that $Z_2^*(\rho, s) = f''(t_0)\, Z^*(\rho/c, s)$, then (2.28) becomes

$$\begin{aligned}\hat{f}_h^{*\prime}(t) = h f''(t_0)\Big[(\hat{h}_{\rm crit}/h)\, Z\big\{\hat{h}_{\rm crit}/(ch_0), (h/\hat{h}_{\rm crit})(s + \hat{s}_h)\big\} + (\hat{s}_h + s) \\ + Z^*\big\{h/(ch_0), s\big\}\Big] + O_p\Big\{\big(h_0^3 \log n\big)^{1/2}\Big\}\,.\end{aligned} \tag{2.29}$$

Furthermore, $\hat{h}_{\text{crit}} = cRh_0 + o_p(h_0)$ and $(\hat{t}_0 - t_0)/\hat{h}_{\text{crit}} = S + o_p(1)$, and so, writing $h = crh_0$,

$$\hat{f}_h^{*\prime}(t) = crh_0 f''(t_0)\{r^{-1} R Z(R, S + R^{-1}rs) + r^{-1}RS + s + Z^*(r,s)\} + o_p(h_0). \tag{2.30}$$

Finally, defining ζ^* and R^* as in Section 2.6.3 completes the proof of Theorem 2.3.

2.7.4 Proof of Theorem 2.4

Let $f_h(x) = \int K(y) f(x - hy)\, dy$, the expected value of $\hat{f}_h$, and take $\mathcal{I}$ to be the whole real line. Silverman (1983) used the variation diminishing properties of the Gaussian kernel and the continuity properties of f_h to show that the number of turning points of f_h in $\mathcal{I}$ is a right-continuous, nonincreasing function of h, for $h \geq 0$. Silverman (1981) proved the same result for $\hat{f}_h$.

Therefore, under the conditions of the theorem, there exist constants $0 < h_1 \leq h_2 < \infty$ such that (i) for all $h < h_1$, f_h has at least two turning points interior to $\mathcal{I}$, and (ii) for all $h > h_2$, f_h has at most one turning point interior to $\mathcal{I}$. Silverman (1983) also proved that as $n \to \infty$ the number of turning points of $\hat{f}_h$, for fixed h, tends to the number of turning points of f_h almost surely. This implies that for each $0 < \epsilon < h_1$,

$$P(h_1 - \epsilon < \hat{h}_{\text{crit}} < h_2 + \epsilon) \to 1$$

as $n \to \infty$.

If $c < h_1$, then $P(\hat{h}_{\text{crit}} > c) \to 1$, which completes the proof of part (a) of the theorem.

Let $\hat{h}^*(h)$ be the version of $\hat{h}^*_{\text{crit}}$ computed if the resample $X_1^*, \ldots, X_n^*$ is drawn from the distribution with density $\hat{f}_h$, instead of from $\hat{f}_{\text{crit}}$. If $h > h_2$ then $\hat{f}_h$ will tend almost surely to a density which has only one turning point interior to $\mathcal{I}$. Hence, conditionally on $\mathcal{X}$ we shall have $\hat{h}^*(h) = O_p(n^{-1/5})$ almost surely, by Theorem 2.2. Hence, for each $0 < \epsilon < h_1$ and $\delta > 0$,

$$\sup_{h_1 - \epsilon < h < h_2 + \epsilon} P\left\{\hat{h}^*(h) \leq \delta \middle| \mathcal{X}\right\} \to 1$$

in probability as $n \to \infty$. Part (b) of the theorem follows from these results.

2.7.5 Proof of Theorem 2.5

Choose $\delta > 0$ so small that the equation $f_h'(x) = f_h''(x) = 0$ has no solutions with $h \in (0, \delta]$ and $x \in \mathcal{I}$. In view of Theorem 2.1, if part (a) of Theorem 2.5 fails then there exists a constant $C_1 > 0$, and sequences of random variables $h(n) \in [C_1 n^{-1/5}, \delta]$ and $x_{h(n)} \in \mathcal{I}$, such that with

$$\mathcal{E}(n) = \left\{ \hat{f}'_{h(n)}(x_{h(n)}) = \hat{f}''_{h(n)}(x_{h(n)}) = 0 \right\},$$

the probability $P\{\mathcal{E}(n)\}$ is bounded away from 0 along an infinite sequence $\{n_k\}$ of n's. (We choose to neglect the third derivative.) We shall show that this leads to a contradiction. Two cases need separate treatment — where $h(n)$ does not converge to zero, and where it does.

Case 1: for some $\epsilon > 0$, $P\{h(n_k) > \epsilon\}$ does not converge to 0. Suppose $P\{h(n_k) > \epsilon\}$ is bounded away from zero. Along a subsequence of values of k, the random vector $(x(n_k), h(n_k))$ converges in distribution. Let (x_1, h_1) be a point of support of its limiting distribution. Necessarily, $h_1 \in [\epsilon, \delta]$. Since $(\hat{f}'_h(x), \hat{f}''_h(x))$ converges to $(f'_h(x), f''_h(x))$, with probability one, uniformly in values of (x, h) in any sufficiently small neighbourhood of (x_1, h_1), then $f'_{h_1}(x_1) = f''_{h_1}(x_1) = 0$, which contradicts our choice of δ. Therefore, Case 1 cannot arise.

Case 2: $h(n_k) \to 0$ in probability. Since $P\{\hat{f}'_{h(n)}(x_{h(n)}) = 0\}$ is bounded away from 0 along a subsequence, and

$$\sup_{C_1 n^{-1/5} \le h \le \delta, x \in \mathcal{I}} \left| \hat{f}'_h(x) - f'_h(x) \right| \to 0$$

in probability; and because $h(n) \to 0$ in probability; then for some $t_0 \in \mathcal{I}$ with $f'(t_0) = 0$, some sequence of constants $\delta_n \to 0$, and all $\epsilon > 0$,

$$\limsup_{n \to \infty} P\left\{ \mathcal{E}(n)\,,\ h(n) \in [C_1 h_0, \delta_n]\,,\ |x_{h(n)} - t_0| \le \epsilon \right\} > 0\,. \tag{2.31}$$

Next we prove the following lemma.

Lemma 2.1 *Assume the conditions of Theorem 3.4, let $t_1, \ldots, t_m$ denote the turning points of f interior to $\mathcal{I}$, put $h_0 = n^{-1/5}$, and let $\delta_n \to 0$ so slowly that $h_0/\delta_n \to 0$. Then*

$$\lim_{C_2 \to \infty} \limsup_{n \to \infty} P\Big\{ \hat{f}'_h(x) = \hat{f}''_h(x) = 0 \text{ for some } h \in [C_2 h_0, \delta_n] \text{ and some } x \in \mathcal{I} \text{ satisfying } \inf_{1 \le k \le m} |x - t_k| > C_2 h_0 \Big\} = 0\,.$$

Proof. Let $t_0 \in \mathcal{I}$ be a turning point of f, and let $j = 1$ or 2. We begin by proving that for all $C_1 > 0$ and $0 < \epsilon < \frac{1}{2}$,

$$E\Big\{ \sup_{u:t_0+\rho h_0 u \in \mathcal{I}} \sup_{C_1 \le \rho \le \delta_n/h_0} (1+|u|)^{-(1/2)-\epsilon}\, \rho^j\, n^{(2-j)/5} \times \big| \hat{f}^{(j)}_{\rho h_0}(t_0+\rho h_0 u) - f^{(j)}_{\rho h_0}(t_0+\rho h_0 u) \big| \Big\} = O(1)\,. \quad (2.32)$$

The approximation of Komlós, Major and Tusnády (1975) gives

$$\hat{f}^{(j)}_h(x) - f^{(j)}_h(x) = \left(n^{1/2} h^{j+1}\right)^{-1} \int K^{(j+1)}(y)\, W^0\{F(x-hy)\}\, dy + R_{1jn}(x,h) \left(nh^{j+1}\right)^{-1} \log n$$

for $n \ge 2$, where

$$E\Big\{ \sup_{x \in \mathcal{I}} \sup_{C_1 h_0 \le h \le \delta_n} |R_{1jn}(x,h)| \Big\} = O(1)\,.$$

Properties of the modulus of continuity of W^0 (e.g. Garsia, 1970) yield, for $n \ge 2$,

$$\int |K^{(j+1)}(y)| \left| W^0\big[F\{t_0+h(u-y)\}\big] - W^0\big\{F(t_0)+h(u-y)\, f(t_0)\big\} \right| dy = R_{2jn}(u,h)\, h\, (\log n)\, (1+|u|)\,,$$

where

$$E\Big\{ \sup_{u:t_0+hu \in \mathcal{I}} \sup_{C_1 h_0 \le h \le \delta_n} |R_{2jn}(u,h)| \Big\} = O(1)\,.$$

Hence, with $\rho = h/h_0$ and

$$W_2(v) = \{h_0\, f(t_0)\}^{-1/2} \left[W^0\{F(t_0) - h_0 v\, f(t_0)\} - W^0\{F(t_0)\} \right]$$

we have

$$\hat{f}^{(j)}_{\rho h_0}(t_0+\rho h_0 u) - f^{(j)}_{\rho h_0}(t_0+\rho h_0 u) = \left(n^{1/2}\, \rho^{j+1}\, h_0^{j+(1/2)}\right)^{-1} f(t_0)^{1/2} \times \int K^{(j+1)}(y)\, W_2\{\rho(y-u)\}\, dy + R_{3jn}(u,\rho) \left(n^{1/2}\, h_0^j\, \rho^j\right)^{-1} \log n\,,$$

where $R_{3jn}(u,h)$ satisfies

$$E\Big\{ \sup_{u:t_0+\rho h_0 u \in \mathcal{I}} \sup_{C_1 \le \rho \le \delta_n/h_0} |R_{kjn}(u,h)| \Big\} = O(1)\,. \quad (2.33)$$

Let $0 < \epsilon < \frac{1}{2}$, and define

$$V_1 = \sup_{|t| \le 1} |W_2(t)|\,, \quad V_2 = \sup_{|t| \ge 1} |t|^{-(1/2)-\epsilon} |W_2(t)| \quad \text{and} \quad V = \max(V_1, V_2)\,.$$

Then $|W_2(t)| \le V\,(1 + |t|)^{(1/2)+\epsilon}$ for all t. Hence, for $j = 1, 2$,

$$\begin{aligned} \int \left| K^{(j+1)}(y)\, W_2\{\rho(y-u)\} \right| dy &\le V \int \left| K^{(j+1)}(y) \right| (1 + \rho|y-u|)^{(1/2)+\epsilon}\, dy \\ &\le C\, V\, (1+\rho)^{(1/2)+\epsilon}\, (1 + |u|)^{(1/2)+\epsilon}\,, \end{aligned}$$

where the constant C depends only on ϵ. Therefore,

$$\begin{aligned} \left| \hat{f}^{(j)}_{\rho h_0}(t_0 + \rho h_0 u) \right. &- \left. f^{(j)}_{\rho h_0}(t_0 + \rho h_0 u) \right| \\ &= R_{4jn}(u, \rho)\, n^{-(2-j)/5}\, \rho^{-j}\, (1 + |u|)^{(1/2)+\epsilon}\,, \end{aligned}$$

where R_{4jn} satisfies (2.33). This implies the desired formula (2.32).

Returning to the proof of Lemma 2.1, we first apply (2.32) for $j = 1$. Given $x \in \mathcal{I}$, let $t_0 = t_0(x) \in \{t_1, \ldots, t_m\}$ denote the turning point to which x is nearest. Since $f''(t_0) \ne 0$ then there exists $C > 0$ such that $|f'_h(x)| > C\,|x - t_0|$ uniformly in values $x \in \mathcal{I}$ that are nearer to t_0 than to any other turning point in $\mathcal{I}$. Hence, with $u = u(x, h) = (x - t_0)/h$,

$$\begin{aligned} |\hat{f}'_h(x)| &\ge |f'_h(x)| - |\hat{f}'_h(x) - f'_h(x)| \\ &\ge C\, h\, |u| - R_{4,1,n}(u, \rho)\, h_0^2\, h^{-1}\, (1 + |u|)^{(1/2)+\epsilon}\,, \end{aligned}$$

and so by (2.32), for all $C_1 > 0$,

$$\lim_{C_3 \to \infty} \limsup_{n \to \infty} P\Big\{ \hat{f}'_h(x) = 0 \text{ for some } h \in [C_1 h_0, \delta_n] \text{ and some } x \in \mathcal{I} \text{ satisfying } \inf_{1 \le k \le m} |x - t_k| > C_3 h \Big\} = 0\,. \tag{2.34}$$

Next we apply (2.32) when $j = 2$. There exists a constant $C > 0$ such that $|f''_h(x)| \ge C$ uniformly in $|x - t_0| \le C_3 h$ and $h \in [C_1 h_0, \delta_n]$, and for such values of (x, h),

$$|\hat{f}''_h(x)| \ge |f''_h(x)| - |\hat{f}''_h(x) - f''_h(x)| \ge C - R_{4,2,n}(u, \rho)\, \rho^{-2}\, (1 + C_3)^{(1/2)+\delta}\,,$$

whence by (2.32) we have for all $C_3 > 0$,

$$\lim_{C_2 \to \infty} \limsup_{n \to \infty} P\Big\{ \hat{f}''_h(x) = 0 \text{ for some } h \in [C_2 h_0, \delta_n] \text{ and some } x \in \mathcal{I} \text{ satisfying } \inf_{1 \le k \le m} |x - t_k| \le C_3 h \Big\} = 0\,. \tag{2.35}$$

Lemma 2.1 follows from (2.34) and (2.35).

In view of Lemma 2.1, result (2.31) will fail if, for all $0 < C_1 < C_2 < \infty$,

$$P\Big\{\hat{f}_h'(x) = \hat{f}_h''(x) = 0 \text{ for some } h \in [C_1 h_0, C_2 h_0] \\ \text{and some } x \in \mathcal{I} \text{ satisfying } |x - t_0| \le C_2 h_0\Big\} \to 0\,.$$

The Wiener–process approximation arguments used during the proof of the lemma may be employed to show that the probability on the left-hand side converges to

$$P\Big\{\int K^{(2)}(y+u)\,W_2(\rho y)\,dy = -\rho^3\,p\,u\,,\ \int K^{(3)}(y+u)\,W_2(\rho y)\,dy = -\rho^3\,p \\ \text{for some } \rho \in [C_1, C_2] \text{ and some } u \text{ satisfying } |u| \le C_2\Big\}\,,$$

where $p = f''(t_0)/f(t_0)^{1/2}$. This probability is zero. Therefore, (2.31) cannot be correct, and so part (a) of Theorem 2.5 must be valid.

In conclusion, we outline derivation of parts (b) and (c) of Theorem 2.5. Suppose f has just m turning points $t_1, \ldots, t_m$ in $\mathcal{I}$. Choose $\epsilon > 0$ so small that $\mathcal{J}_j = [t_j - \epsilon, t_j + \epsilon] \subseteq \mathcal{I}$ and $\mathcal{J}_j$ does not include any t_k's with $k \ne j$; let $h(u) = n^{-1/5}u$; let $M_j(u)$ equal the number of turning points of $\hat{f}_{h(u)}$ in $\mathcal{J}_j$; and put $c_j = f(t_j)^{1/5}/|f''(t_j)|^{2/5}$. The methods employed to derive Theorem 2.2 may be used to show that for arbitrary fixed $0 < C_1 < C_2 < \infty$ the vector $(M_1, \ldots, M_m)$ of processes $M_j(u)$, $u \in [C_1, C_2]$, has a joint weak limit $(L_1, \ldots, L_m)$, where (i) the processes L_j are stochastically independent, and (ii) each L_j has the distribution of the number of zeros of $Y_j(s) = Z(u/c_j, s) + s$. (The processes L_j and M_j are right-continuous step functions with left-hand limits.) The probability that a mode "migrates" away from a turning point converges to 0 as $n \to \infty$. Indeed, if $\delta > 0$ is sufficiently small and $\mathcal{J} \subseteq \mathcal{I}$ denotes a set on which $|f_h'|$ is bounded away from zero uniformly in $0 < h \le \delta$, then, since $\hat{f}_h' - f_h'$ converges to 0 uniformly in $h \in [\epsilon_n n^{-1/5}, \delta]$, provided $\epsilon_n \to 0$ sufficiently slowly, we may show that

$$P\big\{\text{there exists no } x \in \mathcal{J} \text{ or } h \in \big[\epsilon_n n^{-1/5}, \delta\big] \text{ such that } \hat{f}_h'(x) = 0\big\} \to 1\,.$$

In view of these results, parts (b) and (c) of Theorem 2.5 may be derived by making minor changes to the proofs of Theorems 2.2—2.4.

Chapter 3

Numerical properties

3.1 Introduction

This chapter is concerned with numerical aspects of the test for multimodality discussed in Chapter 2. We begin by calibrating the test to make its significance level asymptotically correct. An explicitly defined function for the degree of correction is obtained and we tabulate the size of the adjustment for a large selection of test levels. This is followed by an extensive simulation study of the numerical behaviour of the test for multimodality. We examine the level error for the test and the effect that various adjustments have on this error for a range of unimodal distributions. The chapter concludes with an investigation of the power of the test on bimodal distributions and how this is improved when different corrections are made.

3.2 Calibration

It was explained in Section 2.3 that the problem of calibrating Silverman's test to make its significance level asymptotically accurate is to determine λ_α such that

$$P\Big\{\widehat{G}(\lambda_\alpha) \geq 1-\alpha\Big\} = \alpha,$$

where α is the level of the test. Here $\widehat{G}(\lambda)$ is the stochastic process to which the bootstrap distribution function

$$\widehat{G}_n(\lambda) = P(\hat{h}^*_{\text{crit}}/\hat{h}_{\text{crit}} \leq \lambda | \mathcal{X})$$

weakly converges. The value of λ_α was calculated by simulating the bootstrap test for standard Normal data. Each time the test was performed we obtained a realization of $\widehat{G}_n(\lambda)$. By repeatedly performing the test on different samples drawn from the standard Normal distribution we were able to obtain an approximation to the distribution of $\widehat{G}_n$. Since $\widehat{G}$ is the weak limit of $\widehat{G}_n$, by performing the simulations with n sufficiently large, we could calculate λ_α from our approximation of the distribution of $\widehat{G}_n$.

The precise steps in the calculation of λ_α are described below.

(i) Generate a random sample $\mathcal{X} = \{X_1, \ldots, X_n\}$ from a standard Normal distribution.

(ii) Calculate $\hat{h}_{\text{crit}}$.

(iii) Draw a resample $X_1^*, \ldots, X_n^*$ from $\hat{f}_{\text{crit}}$.

(iv) Calculate $\hat{h}^*_{\text{crit}}$ for the resample.

(v) Repeat steps (iii) and (iv) a large number, say B, of times. The values of $\hat{h}^*_{\text{crit}}/\hat{h}_{\text{crit}}$ can be used as an approximate realization of the distribution function

$$\widehat{G}_n(\lambda) = P(\hat{h}^*_{\text{crit}}/\hat{h}_{\text{crit}} \leq \lambda | \mathcal{X}).$$

(vi) Repeat steps (i) through to (v) K times. This results in K realizations of $\widehat{G}_n(\lambda)$.

(vii) For a predetermined test level α choose λ_α such that

$$P\Big\{\widehat{G}_n(\lambda_\alpha) \geq 1 - \alpha\Big\} = \alpha,$$

where the probability is estimated by the proportion of realizations of $\widehat{G}_n(\lambda)$ greater than $1 - \alpha$.

For the calculation of λ_α we used $B = K = 4999$ and determined that a sample size of $n = 10000$ was adequate. This choice of sample size will be justified later.

Since the Normal distribution has infinite support we restricted the interval on which we were testing for multimodality to the compact interval $\mathcal{I} = [-1.5, 1.5]$. This was done to avoid the problem of detecting spurious modes when testing on

unbounded intervals, which was discussed in Section 2.4. The choice of $\mathcal{I}$ will be studied in Section 3.4.2. To calculate the critical bandwidth $\hat{h}_{\text{crit}}$ (and its bootstrap equivalent $\hat{h}^*_{\text{crit}}$) we constructed kernel density estimates using an algorithm closely based on that of Silverman (1982) and the improvements suggested to it by Jones and Lotwick (1984). This algorithm was used to calculate the value of the density estimate at each point of a grid of 184 evenly spaced points on the interval $[-1.5, 1.5]$. The critical bandwidth was then determined by searching for the smallest bandwidth that produced a density estimate with only one turning point on the interval of interest.

To satisfy ourselves that a grid of 184 points was fine enough we found the critical bandwidth for 1000 different samples from a standard Normal distribution using a grid of 184 points and then repeated the procedure doubling and then quadrupling the number of points. The largest relative change induced in the 1000 critical bandwidths by quadrupling the number of points in the grid was less that 1.5%. There was no difference in the calculated value of λ_α, to 8 significant figures.

We evaluated λ_α using the above algorithm for $\alpha = 0.001, 0.002, \ldots, 0.999$. These values are plotted on Figure 3.1. A function of the form

$$\lambda_\alpha = \frac{a_1\alpha^3 + a_2\alpha^2 + a_3\alpha + a_4}{\alpha^3 + a_5\alpha^2 + a_6\alpha + a_7} \tag{3.1}$$

was fitted to the output to provide a means of approximating λ_α for arbitrary α. The coefficients for this function were fitted using the nonlinear least-squares fitting function (nls) in S-PLUS. These are listed in Table 3.1 and the approximating curve is included on Figure 3.1. Table 3.2 lists the value of the adjustment for a large selection of test levels.

The size of the correction λ_α need not have been calculated by simulating the test on a Normal distribution; any unimodal distribution would have been suitable. We originally attempted the calculation using the distribution with the quadratic density,

$$f(x) = \begin{cases} (3/4)\,(1 - x^2) & \text{if } -1 \le x \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

This density is the so called Epanechnikov kernel function and it was chosen because it had compact support and no tails. However it was discovered that the test behaves a little atypically for samples from this distribution. The test is rather less conservative for data drawn from the Epanechnikov distribution than it is for data

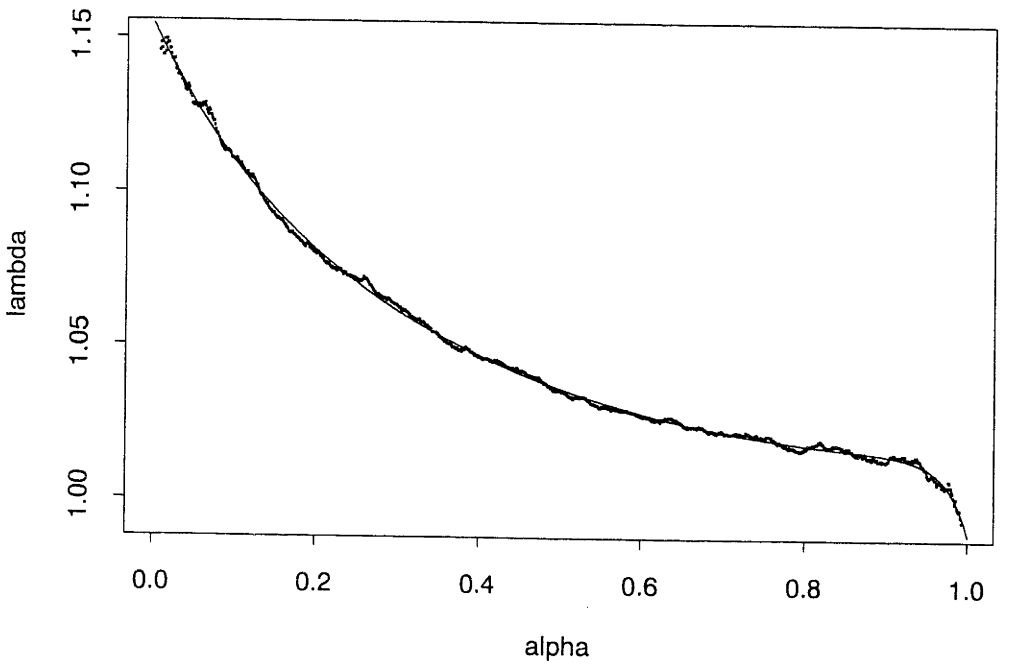


Figure 3.1: *Asymptotic value of λ_α: the dots are the empirically computed values and the line is the approximating function* (3.1).

Table 3.1: *Estimated coefficients for the approximating function given in equation* (3.1).

a_1	a_2	a_3	a_4
0.9402937	-1.5991352	0.1769547	0.4897121

a_5	a_6	a_7
-1.7779310	0.3616166	0.4242250

Table 3.2: *The size of the asymptotic adjustment λ_α for a test of level α.*

α	λ_α	α	λ_α
0.005	1.1516	0.13	1.1001
0.01	1.1489	0.14	1.0971
0.02	1.1436	0.15	1.0942
0.03	1.1387	0.16	1.0914
0.04	1.1339	0.17	1.0887
0.05	1.1294	0.18	1.0861
0.06	1.1252	0.19	1.0837
0.07	1.1211	0.20	1.0813
0.08	1.1172	0.25	1.0705
0.09	1.1135	0.30	1.0615
0.10	1.1099	0.35	1.0537
0.11	1.1065	0.40	1.0471
0.12	1.1032	0.50	1.0364

drawn from all the other distributions which we have considered; see Section 3.6, especially Figure 3.12. We decided to work with the Normal distribution since the test behaves in a similar manner on Normal samples as it does for samples for a wide variety of other distributions. This can be seen from the simulations in Section 3.6.

Alternatively, the calculation of λ_α could have been performed by simulating the stochastic process $\widehat{G}$, the weak limit of $\widehat{G}_n$, which was derived in Section 2.6. This approach presents a number of difficulties, however. While in principle $\widehat{G}$ is known, it depends on two random variables R and R^* whose distributions, particularly that of R^*, are rather complex. It is more straightforward to approximate the asymptotic distributions of $\hat{h}_{\text{crit}}$ and $\hat{h}^*_{\text{crit}}$ by applying the test to large samples. This approach has the added advantages: that it allows us to make use of existing, fast, efficient algorithms for evaluating kernel density estimates, such as those described by Fan and Marron (1994), Jones and Lotwick (1984) and Silverman (1982); and the computing code that is used for calculating λ_α can be reused in simulations for studying the performance of the test.

The drawback of using Monte Carlo methods and simulating the bootstrap test

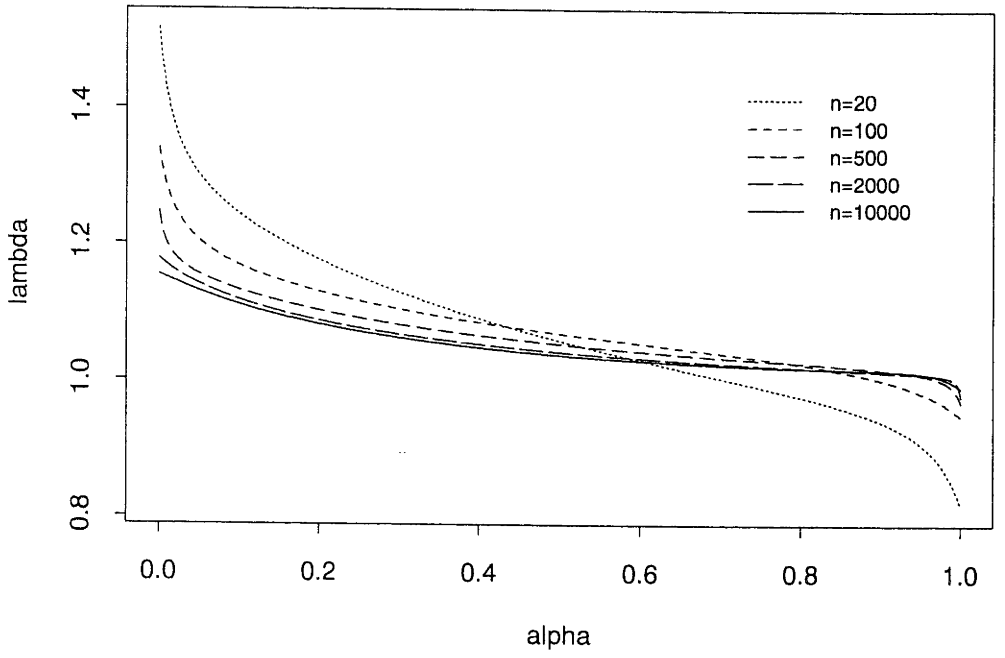


Figure 3.2: *The approximating function* (3.1) *calculated for a range of sample sizes.*

to calculate an asymptotic correction is that we need to determine what sample size is large enough to fully capture the test's asymptotic behaviour. To examine the effect of sample size on λ_α, when testing for multimodality of a Normal distribution, we calculated λ_α for a range of sample sizes: $n = 20$, 50, 100, 200, 500, 1000, 2000, 5000, 10000. Graphs of λ_α against α for a selection of these sample sizes are given in Figure 3.2. It can be observed how the value of λ_α varies with sample size and that it becomes stationary around the $n = 5000$ to 10000 mark. This is more clearly seen in Figure 3.3. These plots reveal the relationship between λ_α and sample size for the commonly used test levels: $\alpha = 0.01$, 0.05, 0.1, 0.2. On the basis of these results we decided that we could use a sample size of $n = 10000$ to adequately describe the asymptotic behaviour of the test.

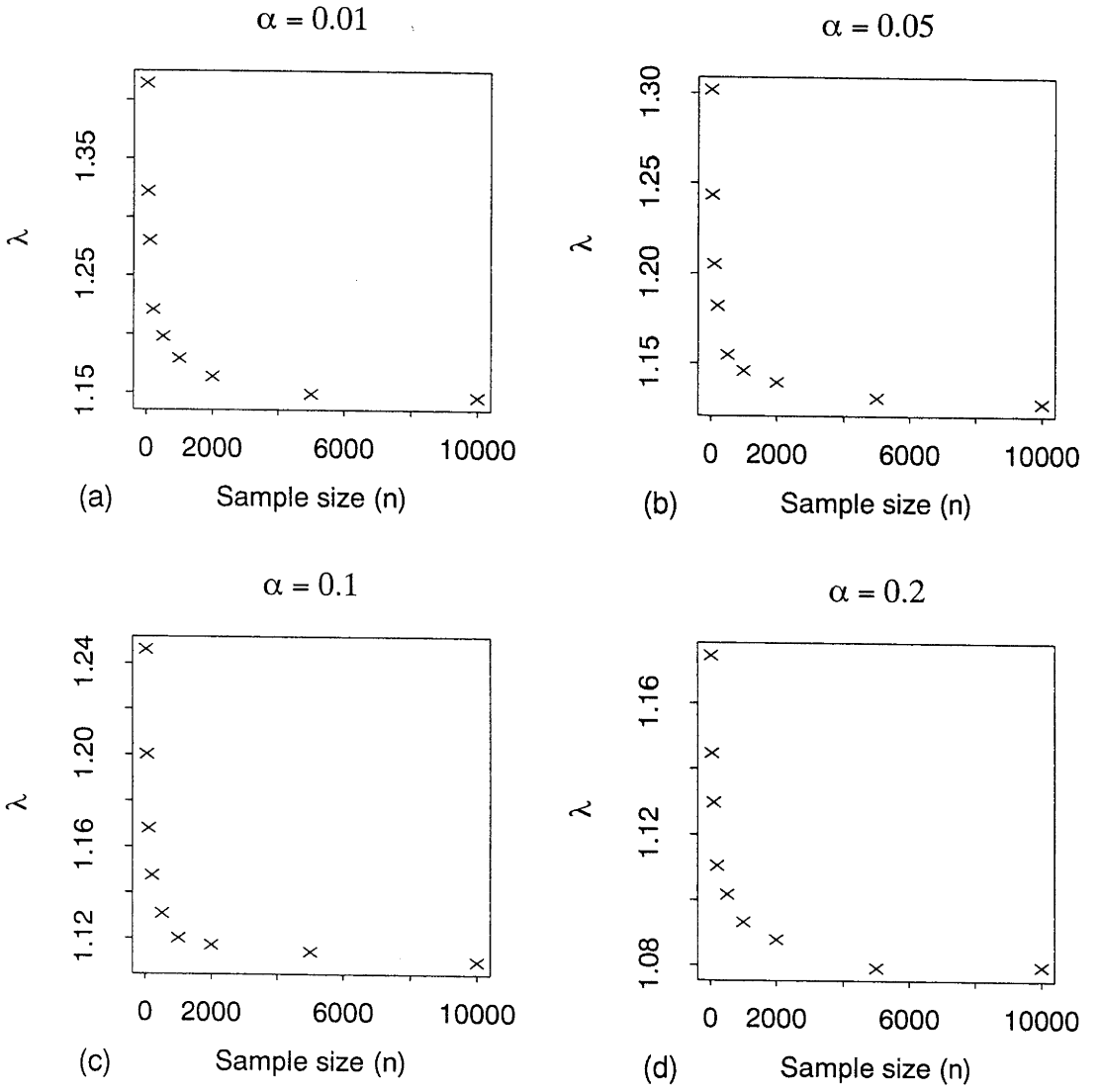


Figure 3.3: *The convergence of λ_α as a function of sample size for (a) $\alpha = 0.01$; (b) $\alpha = 0.05$; (c) $\alpha = 0.1$ and (d) $\alpha = 0.2$.*

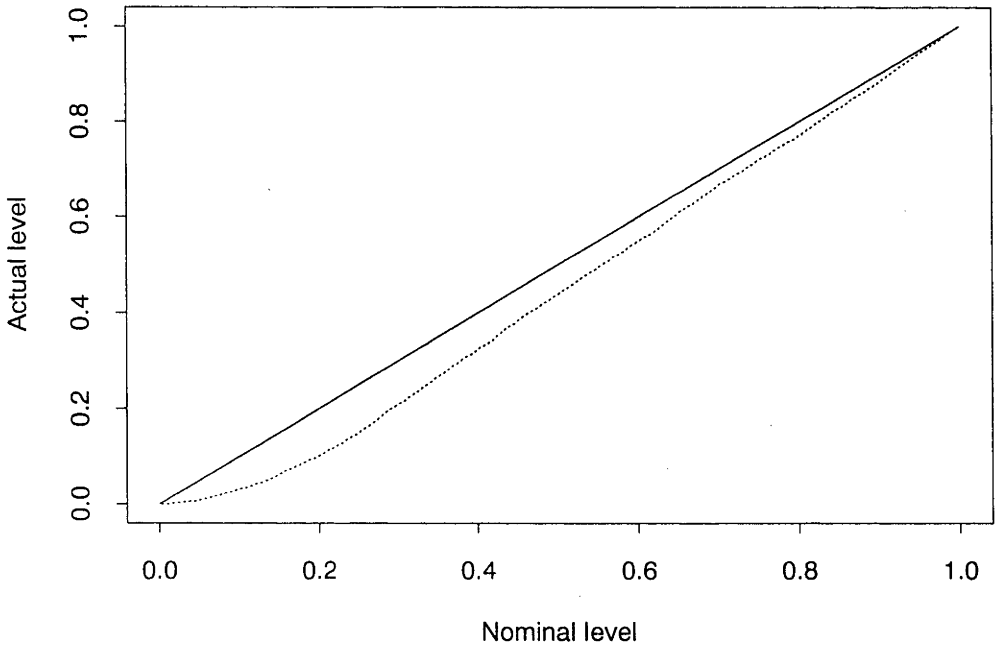


Figure 3.4: *The asymptotic level of Silverman's critical bandwidth test.*

3.3 Large-sample behaviour of the test

The critical bandwidth test is based on the idea that large values of $\hat{h}_{\text{crit}}$ are associated with multimodality, since a considerable amount of smoothing will be required to yield a unimodal density estimate. Therefore we wish to reject values of $\hat{h}_{\text{crit}}$ that are too large to be consistent with the null hypothesis that f is unimodal. The bootstrap test provides a means of gauging the size of $\hat{h}_{\text{crit}}$. However, rejecting H_0 when

$$P\left(\hat{h}^*_{\text{crit}}/\hat{h}_{\text{crit}} \leq \lambda_\alpha \middle| \mathcal{X}\right) \geq 1 - \alpha$$

leads to a test which is asymptotically conservative when conducted, as it usually is, with $\lambda_\alpha \equiv 1$.

By conservative we mean that the actual level of the test is smaller than the nominal level α. This means that the test is less likely to reject the null hypothesis than we expect, which results in a test with reduced power. In the previous section

we calculated the value of λ_α that produces a test with asymptotically correct level α. The actual level of the test obtained when λ_α is replaced by 1 is listed in Table 3.3 and in Figure 3.4 is plotted against the nominal level of the test. The level was computed in the course of the simulations that were performed to calculate λ_α. See the previous section for the computational details. It can be seen from Table 3.3 that the test is extremely conservative. For a nominal level of $\alpha = 0.05$ the actual level is only 0.010, for $\alpha = 0.1$ it is 0.032 and for $\alpha = 0.2$ it is 0.102. Table 3.4 gives the nominal levels that correspond to standard actual levels.

The level of the test is incorrect because the use of $\lambda_\alpha \equiv 1$ is based on the notion that the distribution of

$$U_n = P\big(\hat{h}^*_{\text{crit}} \le \hat{h}_{\text{crit}} \big| \mathcal{X}\big)$$

is close to being Uniform on the interval $(0, 1)$, at least for large values of n. In Sections 2.6 and 2.7 this was shown to be untrue since $\hat{h}_{\text{crit}}$ and $\hat{h}^*_{\text{crit}}$ have quite different limiting distributions. There it was proven that, under H_0, $\hat{h}_{\text{crit}}$ converged in distribution to

$$\big\{n^{-1} f(t_0)/|f''(t_0)|^2\big\}^{1/5} R, \tag{3.2}$$

where t_0 is the location of the mode of f and R is a random variable whose distribution does not depend on the (unknown) density f. The value of f'' at the mode plays a vital role in the distribution of $\hat{h}_{\text{crit}}$; the flatter the mode the greater the value of $\hat{h}_{\text{crit}}$.

The bootstrap test fails to capture the first-order asymptotics of equation (3.2) (Hall and Wood, 1996). This is because $\hat{f}''_{\text{crit}}$ does not consistently estimate f'' since $\hat{h}_{\text{crit}}$ is of order $n^{-1/5}$. It is well known that for bandwidths of this size, $\hat{f}''_h$ does not consistently estimate f'' (see Wand and Jones, 1995, Section 2.12). Bandwidths of a larger order are required for the variance of $\hat{f}''_h$ to tend to zero. This inconsistency results in $\hat{h}^*_{\text{crit}}$ having a different limiting distribution to $\hat{h}_{\text{crit}}$. The asymptotic distribution of $\hat{h}^*_{\text{crit}}$, conditional on $\mathcal{X}$, is

$$\big\{n^{-1} f(t_0)/|f''(t_0)|^2\big\}^{1/5} R^*,$$

where R^* is a random variable whose distribution does not depend on f but it does depend on $\hat{h}_{\text{crit}}$ and $\hat{t}_0$, the location of the mode of $\hat{f}_{\text{crit}}$.

The failure of the bootstrap to consistently estimate the distribution of $\hat{h}_{\text{crit}}$ is demonstrated in Figure 3.5 (a). This graph plots $P\big(\hat{h}^*_{\text{crit}} \le \hat{h}_{\text{crit}} \big| \mathcal{X}\big)$, the bootstrap

Table 3.3: *The asymptotic level of Silverman's test.*

Nominal level	Actual level	Nominal level	Actual level
0.005	0.000	0.130	0.050
0.010	0.000	0.140	0.057
0.020	0.002	0.150	0.062
0.030	0.004	0.160	0.070
0.040	0.006	0.170	0.079
0.050	0.010	0.180	0.088
0.060	0.012	0.190	0.094
0.070	0.016	0.200	0.102
0.080	0.021	0.250	0.149
0.090	0.025	0.300	0.202
0.100	0.032	0.350	0.252
0.110	0.038	0.400	0.308
0.120	0.043	0.500	0.423

Table 3.4: *The nominal levels corresponding to standard actual levels.*

Nominal level	Actual level
0.050	0.010
0.130	0.050
0.198	0.100
0.297	0.200

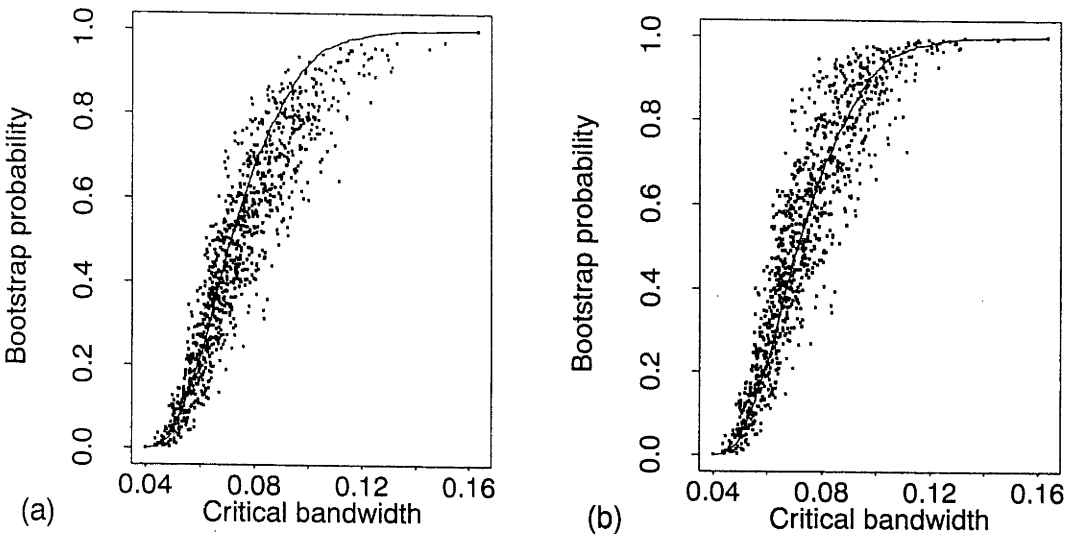


Figure 3.5: *Scatter plot of* $P(\hat{h}^*_{\text{crit}} \leq \hat{h}_{\text{crit}} | \mathcal{X})$ *against* $\hat{h}_{\text{crit}}$ *for 999 simulations of the bandwidth test on standard Normal samples of size* $n = 50000$. *The line is the empirical distribution function of* $\hat{h}_{\text{crit}}$. *Panel (a) is for the uncalibrated form of the test and panel (b) is for the calibrated test.*

estimate of the distribution of $\hat{h}_{\text{crit}}$, against $\hat{h}_{\text{crit}}$ and is superimposed by the empirical distribution function of $\hat{h}_{\text{crit}}$. Each point on the graph represents one of 999 simulations of the test on samples of size $n = 50000$, drawn from a standard Normal distribution. Each bootstrap probability is estimated by calculating $\hat{h}^*_{\text{crit}}$ for 999 resamples. The two most noteworthy features of this graph are: the bootstrap probabilities are quite variable; and as $\hat{h}_{\text{crit}}$ increases the bootstrap probabilities become increasingly biased.

The variability of the bootstrap probabilities is caused by the variance of $\hat{f}''_{\text{crit}}$ not tending to zero. Mammen, Marron and Fisher (1992) conjecture that the bias is caused by the flatness of the mode of $\hat{f}_{\text{crit}}$. Their reasoning is that $\hat{f}'_{\text{crit}}$ vanishes twice in the vicinity of t_0; once at the mode and once at the shoulder. Hence, $\hat{f}''_{\text{crit}}$ tends to be small in this neighbourhood. Consequently, $\hat{h}^*_{\text{crit}}$ will tend to be larger than $\hat{h}_{\text{crit}}$ since $\hat{f}_{\text{crit}}$ will tend to be flatter about its mode than f. This problem is more severe for larger values of $\hat{h}_{\text{crit}}$ because these result in a flatter $\hat{f}_{\text{crit}}$. The density estimate in Figure 3.6 is a typical example of the appearance of $\hat{f}_{\text{crit}}$ for a large value of $\hat{h}_{\text{crit}}$. As a result, asymptotically, $P(\hat{h}^*_{\text{crit}} \leq \hat{h}_{\text{crit}} | \mathcal{X})$ will underestimate

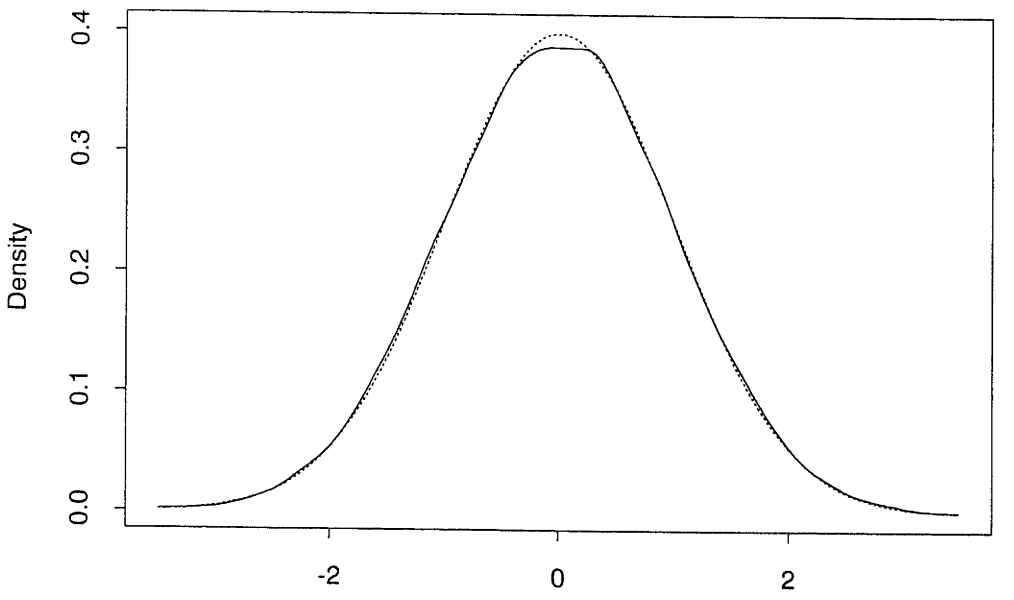


Figure 3.6: *A typical example of* $\hat{f}_{\text{crit}}$ *for a large value of* $\hat{h}_{\text{crit}}$ *(solid line). The dotted line is the standard Normal density.*

the distribution of $\hat{h}_{\text{crit}}$ and

$$P\Big\{P\big(\hat{h}^*_{\text{crit}} \le \hat{h}_{\text{crit}} \big| \mathcal{X}\big) \ge 1-\alpha\Big\} \le \alpha,$$

at least for reasonably small α. Therefore, the test will be conservative.

In fact, Figure 3.4 indicates that asymptotically the test is conservative for all $0 < \alpha < 1$. Unfortunately the level error is worst for $\alpha < 0.25$, which is that range of levels at which a hypothesis test would normally be conducted. However, our calibrated method produces a test with correct level, asymptotically, and the effect of it on the bootstrap probabilities can be seen in Figure 3.5 (b).

3.4 Behaviour of the test on Normally distributed samples of finite size

We have seen in the previous chapter that the use of $\lambda_\alpha \equiv 1$ in the critical bandwidth test results in a conservative test with a substantial asymptotic level error. In this section we shall study how the level of the bandwidth test varies with sample size n and choice of the interval $\mathcal{I}$ for finite sized Normal samples. Throughout this section we shall use the version of Silverman's test without any correction for the variance inflation of the kernel density estimator. Since we are principally interested in how level changes with n and $\mathcal{I}$, rather than the level itself, we have chosen to work with the simplest form of the test. The variance correction will improve the level accuracy of the test but it will not significantly affect the influence of n and $\mathcal{I}$ on level. In the next section we will examine the use of correcting the variance of the resampled data, and other approaches, for reducing the conservatism of the test.

3.4.1 Effect of sample size

To examine the influence of sample size on the performance of the test we simulated the test on samples of size $n = 20$. This was done by drawing samples from a standard Normal distribution and performing the test, over the interval $\mathcal{I} = [-1.5, 1.5]$, on each sample. The results are displayed in Figure 3.7. The top-left panel graphs the actual level of the test against the nominal level, α. If we compare this graph with its large-sample version, Figure 3.4, we can see that for $n = 20$ the test is much more conservative for $\alpha < 0.5$, while for $\alpha > 0.8$ the actual level is greater than the nominal one.

These features are explained by the plot of the empirical distribution of the 999 calculated values of $\hat{h}_{\text{crit}}$ and their corresponding bootstrap probabilities $P(\hat{h}^*_{\text{crit}} \leq \hat{h}_{\text{crit}} | \mathcal{X})$ in Figure 3.7 (b). For smaller values of $\hat{h}_{\text{crit}}$ the bootstrap probabilities lie above the empirical distribution while for the larger values of $\hat{h}_{\text{crit}}$ the bootstrap probabilities lie well below it. The small values of $\hat{h}_{\text{crit}}$ tend to be associated with samples where most of the data points are clustered close to zero, or are far enough outside the interval $\mathcal{I}$ not to affect the number of modes in $\mathcal{I}$. These samples yield an $\hat{f}_{\text{crit}}$ that is tall and peaked and has a shoulder away from the mode; see Figure 3.7 (c). Resampling from such a density will yield values of $\hat{h}^*_{\text{crit}}$ that will usually be smaller than the values of $\hat{h}_{\text{crit}}$ that would be obtained from sampling

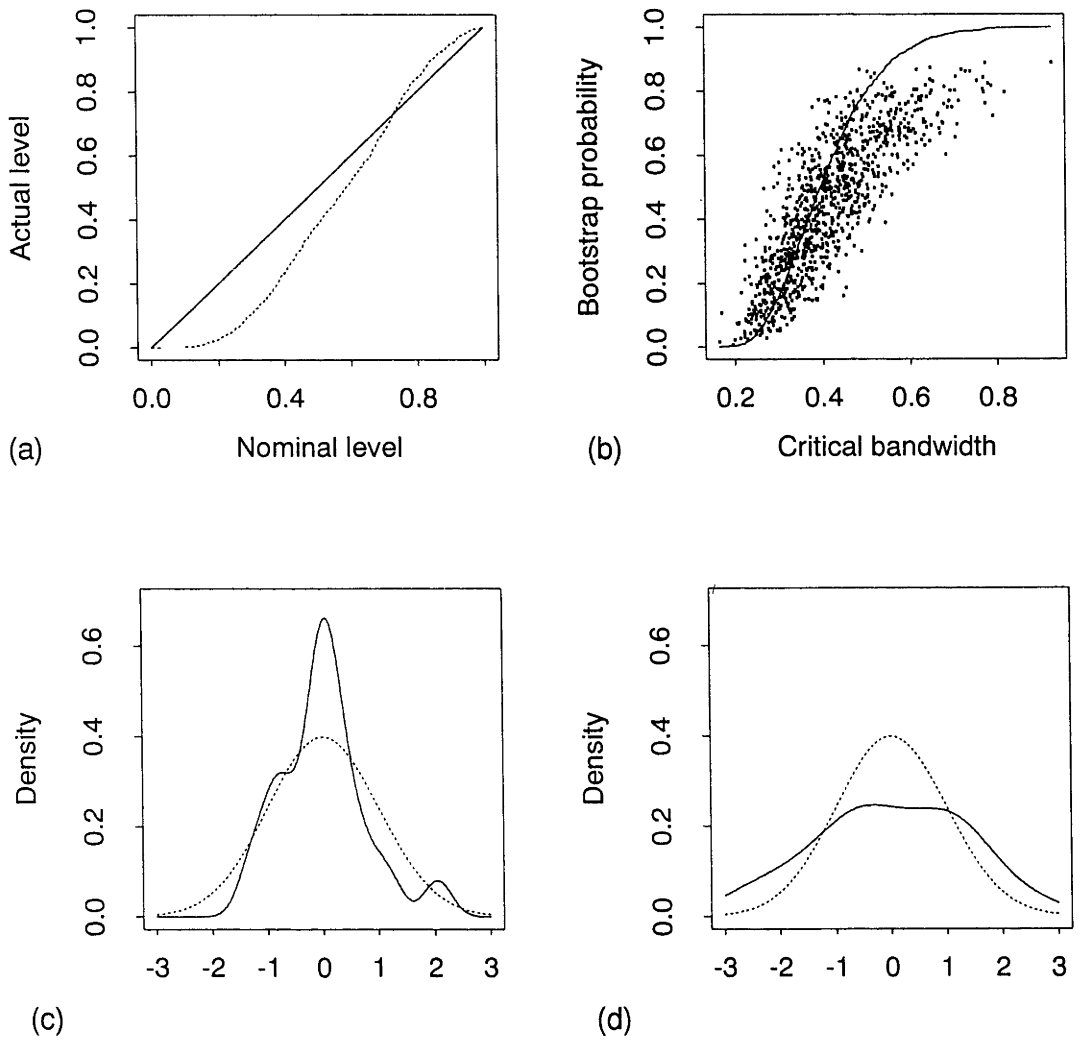


Figure 3.7: *Simulation results for a Normal sample of size* $n = 20$. *Panel (a) gives the usual plot of level accuracy for the bandwidth test. Panel (b) is the plot of* $P(\hat{h}^*_{\text{crit}} \leq \hat{h}_{\text{crit}} | \mathcal{X})$ *against* $\hat{h}_{\text{crit}}$. *Panels (c) and (d) are examples of* $\hat{f}_{\text{crit}}$. *The dotted lines are the standard Normal density.*

from a standard Normal distribution. Hence, in this case, the bootstrap distribution overestimates the distribution of $\hat{h}_{\text{crit}}$.

This situation only occurs with significant frequency for small sample sizes because it becomes increasingly rare, as the sample size increases, that one would obtain a sample where nearly all the observations are clustered very close to the mode. In fact this feature is practically non-existent by the time the sample size reaches $n = 100$. This simulation shows that while the critical bandwidth test appears to be conservative for all α-levels asymptotically, this is not true for quite small sample sizes.

However, for $\alpha \leq 0.5$ the test is extremely conservative. The conservatism of the test for $n = 20$ is caused by the flatness of the resampling density $\hat{f}_{\text{crit}}$. This was the case for the asymptotic situation as well, but for $n = 20$ the effects are far greater. When $\hat{h}_{\text{crit}}$ is large the corresponding $\hat{f}_{\text{crit}}$ tends to be very flat about its mode, like the density pictured in Figure 3.7 (d). Resampling from a flat distribution leads to larger values of $\hat{h}^*_{\text{crit}}$ than would be obtained from sampling from a standard Normal distribution. Consequently, the bootstrap distribution severely underestimates the distribution of $\hat{h}_{\text{crit}}$, resulting in an extremely conservative test.

These simulations were repeated for samples of size $n = 100$ and $n = 1000$. The actual level of the test is plotted against the nominal level in Figure 3.8 and the levels for $n = 20$ and $n = 10000$ are also included. The main point of interest is that the size of the level error decreases, for $\alpha < 0.5$, as the sample size increases. We can also see that as the sample size increases the observed level of the test falls below the nominal level of the test, α, for $\alpha > 0.8$. By the time the sample size reaches $n = 1000$ the test is performing very closely to its asymptotic behaviour; its level error is reasonably similar to the asymptotic one and the test appears to be conservative for all α-levels.

3.4.2 Effect of the choice of the interval $\mathcal{I}$

Up to this point we have only been examining the performance of the test when it is conducted over the interval $\mathcal{I} = [-1.5, 1.5]$. We have discussed in Section 2.4 the necessity of conducting the test over some compact interval $\mathcal{I}$ when the density f, from which the data are drawn, has unbounded support. If we conduct the test over the infinite support of the density then the behaviour of the test will be driven by the spacings between extreme observations in the data set. This will mean that the

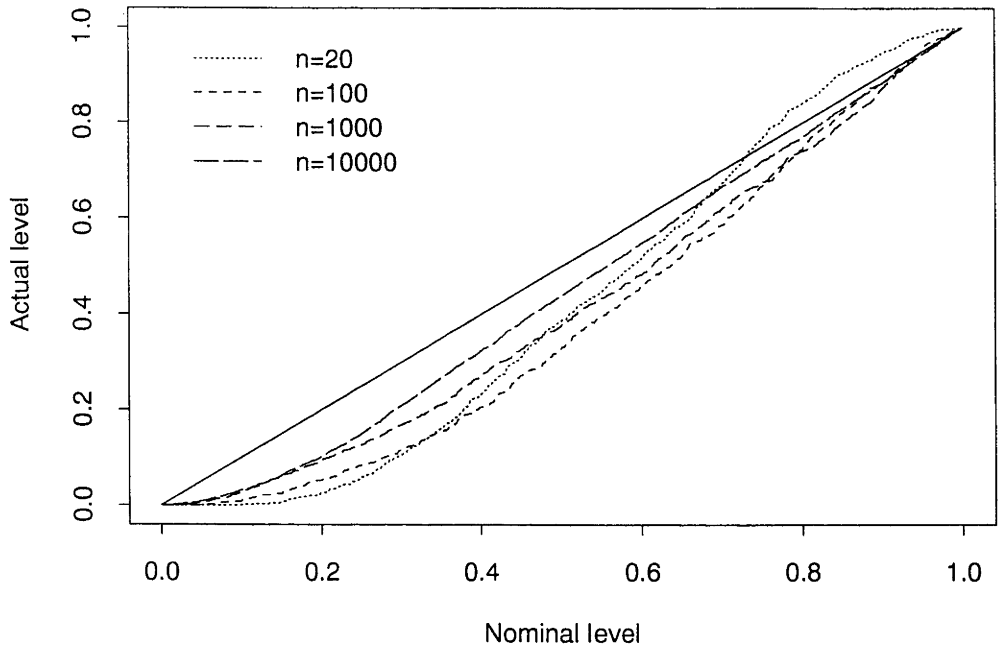


Figure 3.8: *Level accuracy for Normal samples of a range of sizes.*

value of $\hat{h}_{\text{crit}}$ will reflect the behaviour of the tails of f rather than its modality.

Obviously, as the length of the interval $\mathcal{I}$ increases the value of $\hat{h}_{\text{crit}}$ cannot decrease. What is not so obvious is what will happen to the size of the test's p-value; it is unclear how the distribution of $\hat{h}^*_{\text{crit}}$ will be affected in relation to $\hat{h}_{\text{crit}}$. In order to examine the effect of varying the interval $\mathcal{I}$ we performed Silverman's test over the intervals $[-1.5, 1.5]$, $[-2.5, 2.5]$, $[-3.5, 3.5]$ and $[-4.5, 4.5]$ for a wide range of sample sizes, drawn from a standard Normal distribution. The general pattern of the interaction between sample size and interval length can be seen by just considering the simulation results for the two sample sizes, $n = 20$ and $n = 1000$, for the two intervals, $[-1.5, 1.5]$ and $[-3.5, 3.5]$.

The results of these simulations are summarised in Figure 3.9. The top-left graph is the plot of the observed level of the test against the nominal level for the two intervals, for a sample of size $n = 20$. The test is more conservative on the shorter interval, $[-1.5, 1.5]$ (the dotted line) than on the interval $[-3.5, 3.5]$ (the dashed

line) for nominal levels of less than 0.2, and the opposite is true for greater levels. The levels plotted here are for the uncalibrated version of Silverman's test without any correction for the variance inflation of the kernel density estimate. For other versions of the test the position of these two curves will alter but the relationship between the two will essentially remain the same.

The bottom-left panel is a kernel density estimate (with $h = 0.05$) of the increase in p-value recorded when the interval is lengthened from $[-1.5, 1.5]$ to $[-3.5, 3.5]$. (Negative values indicate a decrease in p-value.) Recall that for the uncalibrated version of the test the p-value is

$$p = P(\hat{h}_{\text{crit}} \leq \hat{h}^*_{\text{crit}} \mid \mathcal{X}).$$

When the interval is lengthened one of two things can happen:

(a) $\hat{h}_{\text{crit}}$ can remain the same; or

(b) $\hat{h}_{\text{crit}}$ can increase.

In case (a) the values of $\hat{h}^*_{\text{crit}}$ cannot decrease and therefore the p-value can only increase. In case (b) the p-value may increase or decrease. On the plot we have decomposed the density into these two components. The dashed curve is the component corresponding to case (a) and the dotted curve corresponds to case (b). (Clearly for case (a) the support of the density should not contain any positive values. For the estimate shown this is not the case. Since this density plot is only intended for illustrative purposes we have not used a boundary kernel or some other type of boundary correction.)

The most interesting feature of this density is the long tail on the left-hand side. Since large values of $\hat{h}_{\text{crit}}$ tend to be associated with small p-values and smaller values of $\hat{h}_{\text{crit}}$ with larger p-values, these large decreases in p-value tend to be associated with large increases in $\hat{h}_{\text{crit}}$. These large increases in $\hat{h}_{\text{crit}}$ are caused by the increased amount of smoothing that is required to smooth away any outlying modes in the intervals $[-3.5, 1.5)$ and $(1.5, 3.5]$.

These changes in the p-values explain the differences between the two curves in the top-left panel. We earlier observed that for nominal levels of less than 0.2 the test is less conservative on the larger interval. This is due to the long left tail in the bottom plot. These large decreases in the p-value result in more small p-values when the length of the interval is increased. However for nominal levels greater than

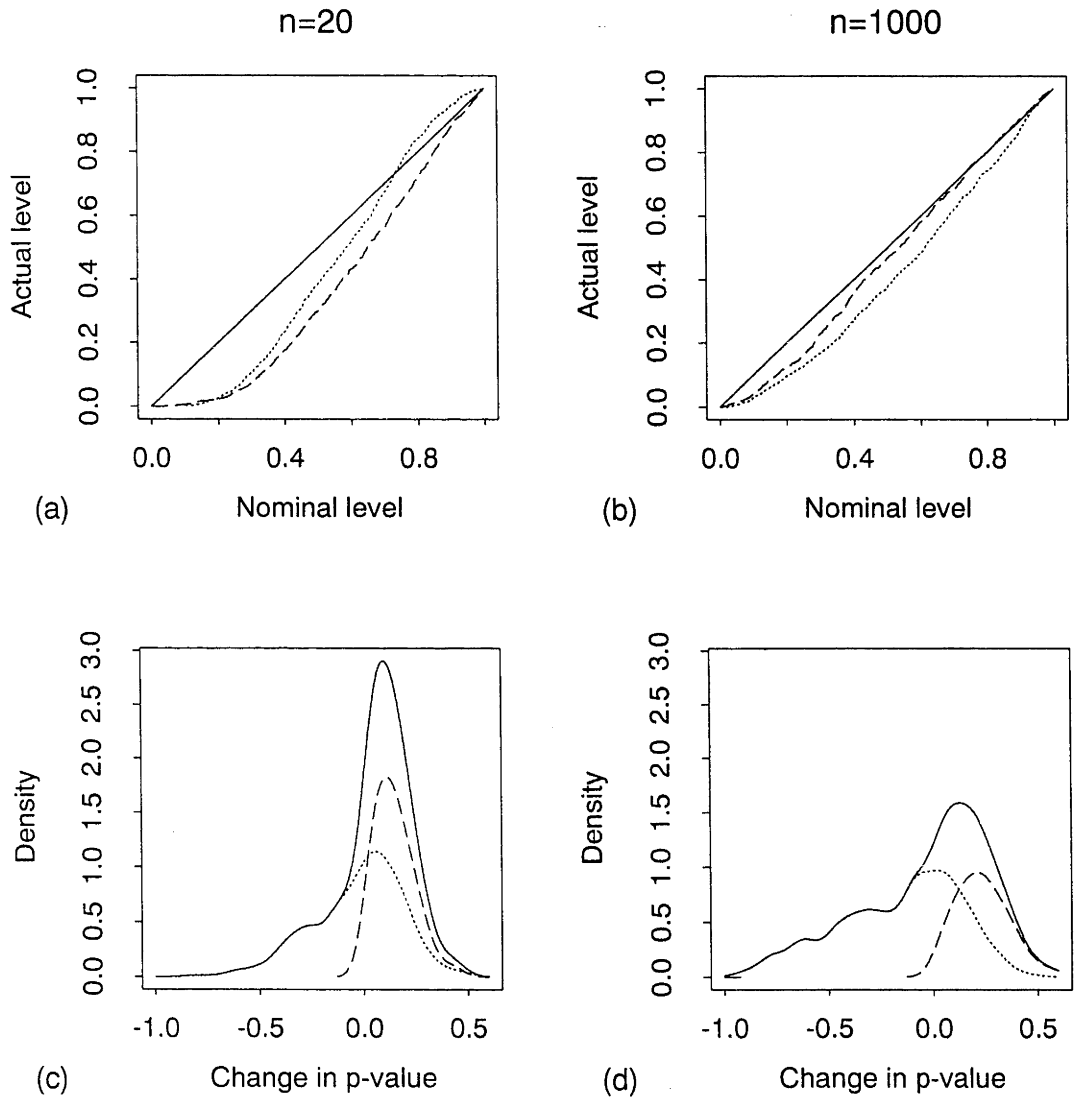


Figure 3.9: *Panel (a) gives the plot of level accuracy for a Normal sample of size* $n = 20$ *for the two intervals* $\mathcal{I} = [-1.5, 1.5]$ *(dotted line) and* $\mathcal{I} = [-3.5, 3.5]$ *(dashed line). Panel (b) is the same as (a) but for* $n = 1000$*. Panel (c) is a kernel density estimate* $(h = 0.05)$ *of the increase in p-value induced by lengthening the interval from* $[-1.5, 1.5]$ *to* $[-3.5, 3.5]$*, for* $n = 20$*. Panel (d) is the same for* $n = 1000$*.*

0.2, the test becomes more conservative as the interval lengthens. This is due to the general increase in p-values; in our simulations the p-value increased as the interval lengthened for 735 out of the 999 data sets that were generated. This large number of increases was primarily due to the frequency with which case (a) occurs; in our simulations this occurred for 460 of the 999 data sets.

The right-hand side of Figure 3.9 contains the same two plots as the other side but for a sample size of $n = 1000$. Here we see that increasing the length of the interval makes the test less conservative for all nominal levels. The main differences from the $n = 20$ setting are that case (a) occurs less frequently – 334 out of 999 times – and the left-hand tail of the density in the bottom-right graph is much heavier.

For this larger sample size there are more observations in the extreme regions $[-3.5, -1.5)$ and $(1.5, 3.5]$. The increased amount of data in these regions means that case (a) will occur less frequently since a greater amount of smoothing will often be required to smooth away modes in these intervals. This increase in the value of $\hat{h}_{\text{crit}}$ will cause a corresponding decrease in the p-values; hence the heavy left-hand tail. Overall there is less of a general increase in p-values for $n = 1000$ than for $n = 20$. Consequently the p-value decreases more often for $n = 1000$ and, when it does decrease, the decrease tends to be greater. This can be seen from the heavier left-hand tail for $n = 1000$ than for $n = 20$, and is the reason that the test becomes less conservative for a sample size of $n = 1000$ when the interval $\mathcal{I}$ is lengthened.

While there is a reasonably complex interaction between the sample size and the length of the interval over which we are testing, it is reassuring to observe that for nominal levels of less than 0.2 the actual level of the test is not particularly sensitive to choice of the interval for either of the two different sample sizes. For $n = 20$ the greatest difference in actual level for nominal levels of less than 0.2 is 0.010, and for $n = 1000$ the greatest difference in level is 0.032. These differences are relatively small, particularly when one notes that the interval $[-3.5, 3.5]$ is over twice the length of $[-1.5, 1.5]$.

3.5 Corrections to reduce conservatism

In the previous section we saw how extremely conservative Silverman's test is, and how this conservatism varies with sample size and the length of the interval over

which we are testing for multimodality. We will now examine the performance of techniques for correcting the test of this conservatism. In Section 1.3 we explained that if our resample $X_1^*, X_2^*, \ldots, X_n^*$ is drawn from the distribution with density $\hat{f}_{\text{crit}}$, then our resampled data will have a variance of $\sigma^2 + \hat{h}_{\text{crit}}^2$, where σ^2 is the variance of the original data. One solution to this is to rescale the resampled data so that they have the same sample mean and variance as the original data; see equation (1.6). By reducing the spread of the resample we shall also reduce the values of $\hat{h}_{\text{crit}}^*$ obtained for the resamples, thus making the test less conservative. During the discussion of the following simulation results we shall refer to this adjustment as the *variance correction*.

In Chapter 2 we formulated a calibrated version of the test based on the limiting distribution of the bootstrap distribution function,

$$\widehat{G}_n(\lambda) = P\big(\hat{h}_{\text{crit}}^*/\hat{h}_{\text{crit}} \leq \lambda \big| \mathcal{X}\big)\,.$$

This limiting distribution does not depend on the unknown density f and the percentage points of this distribution were calculated numerically in Section 3.2. We will refer to this form of the test as the *asymptotic calibration*. In Section 2.3 another type of calibration, based on Monte Carlo techniques, was discussed. We do not consider it here but will assess its performance in Section 3.7.

The variance correction is a small-sample adjustment; for small samples $\hat{h}_{\text{crit}}$ will be relatively large and the effect of correcting the variance from $\sigma^2 + \hat{h}_{\text{crit}}^2$ to σ^2 will be quite marked. On the other hand, for large samples $\hat{h}_{\text{crit}}$ tends to zero and hence the effect of this variance inflation will be negligible; it has no effect on the asymptotic conservatism of the test. Therefore we expect that for large samples the variance correction will do little towards improving the level accuracy of the test. For smaller samples the variance correction does make very significant improvements to the level accuracy, but it is well known (Fisher, Mammen and Marron, 1994) that even with the variance correction the test is still very conservative.

In contrast, the asymptotic calibration is designed for large samples. The theoretical results derived in Chapter 2 show that for very large samples the asymptotic calibration will produce a test with virtually correct level. However, the simulations of the previous section reveal that, at least for the Normal distribution, the test is at its least conservative asymptotically. Hence, an asymptotic calibration will undercorrect the level error for smaller samples, thus producing a test which is still conservative.

The discussion of the previous two paragraphs suggests that by combining the variance correction and the asymptotic calibration we might produce a test with good level accuracy for finite sample sizes. In the following simulation study we compare the performance of four forms of Silverman's test. They are:

(a) no variance correction and no calibration;

(b) variance correction with no calibration;

(c) no variance correction but with asymptotic calibration; and

(d) variance correction with asymptotic calibration.

We begin our simulation study by investigating the behaviour of these four forms of the test when they are carried out on the standard Normal distribution. For each of the sample sizes $n = 50$, $n = 100$ and $n = 200$ we generated 1000 realizations of the sample $\mathcal{X}$. For each of these samples 999 resamples were drawn to approximate the bootstrap distribution function,

$$\widehat{G}_n(\lambda) = P\big(\hat{h}^*_{\text{crit}}/\hat{h}_{\text{crit}} \leq \lambda \big| \mathcal{X}\big)\,.$$

For each method the actual probabilities of rejecting the null hypothesis were approximated by the proportion of times the null hypothesis was rejected. The results are presented in Figure 3.10.

In Figure 3.10 (a) there is a plot of the sampling distribution, the standard Normal. Since the Normal has unbounded support it is necessary to conduct the test over some compact interval $\mathcal{I}$. We chose $\mathcal{I}$ to be the interval $[-1.5, 1.5]$. Earlier, in Section 3.4.2 we saw that the performance of the test is not particularly sensitive to the choice of interval. In each of the other three panels of Figure 3.10 there are graphs of the actual level of the test against the nominal level, for the four methods. Each panel contains this plot for samples of different sizes. In all three panels it can be seen that form (d) of the test consistently performs the best, producing a test whose actual level is extremely close to the correct one. As expected, form (a) is the most conservative but the level error does decrease as the sample size increases for $n = 50$ to $n = 200$. For $n = 50$ forms (b) and (c) perform almost identically but as the sample size increases form (b) (the finite sample variance correction) starts to be outperformed by form (c) (the asymptotic calibration).

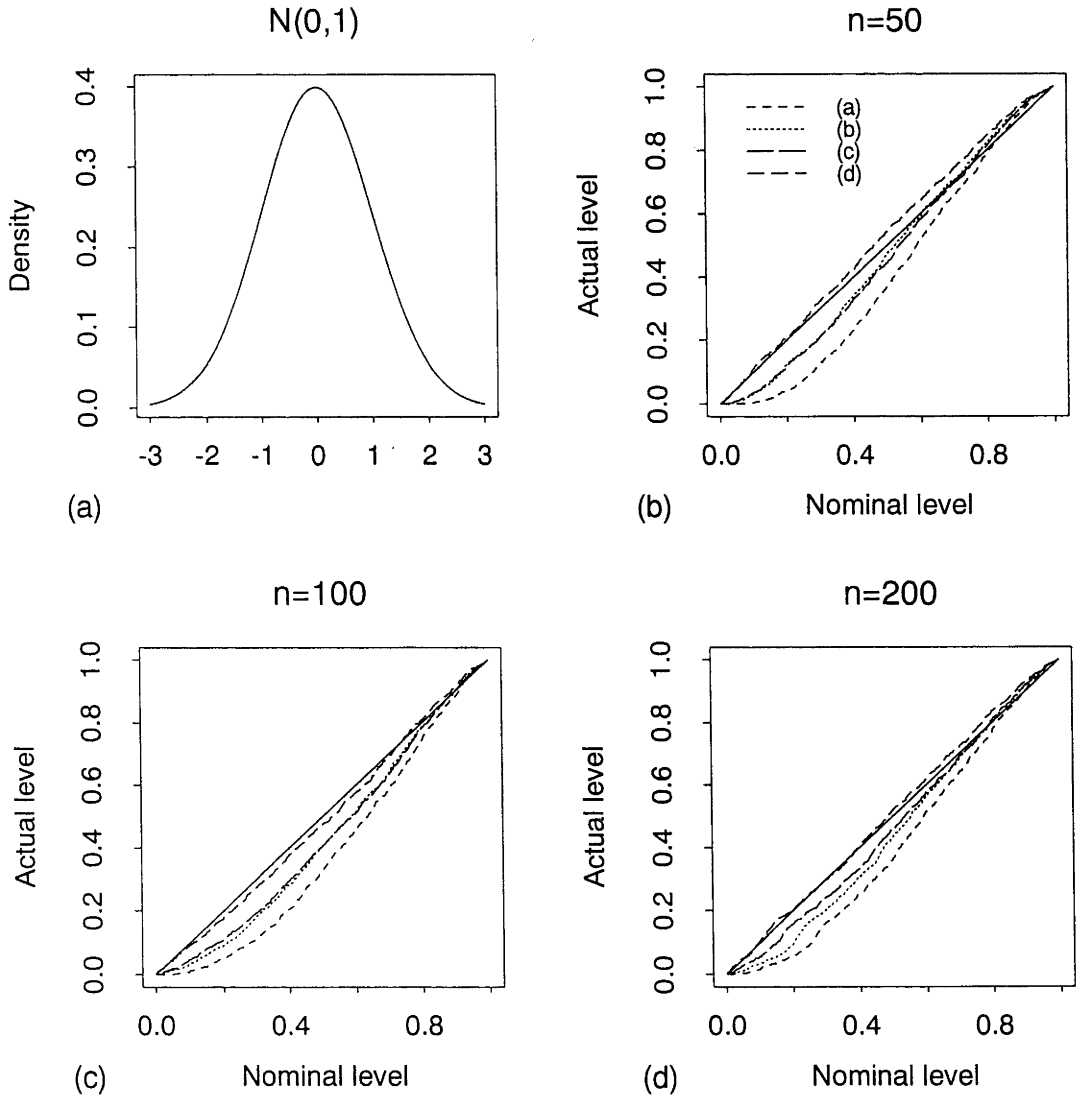


Figure 3.10: *Plot of the level accuracy of forms (a)-(d) of the test. Panel (a) is the sampling distribution, the standard Normal. Panels (b)-(d) give the level accuracy for samples of size* $n = 50$, $n = 100$ *and* $n = 200$.

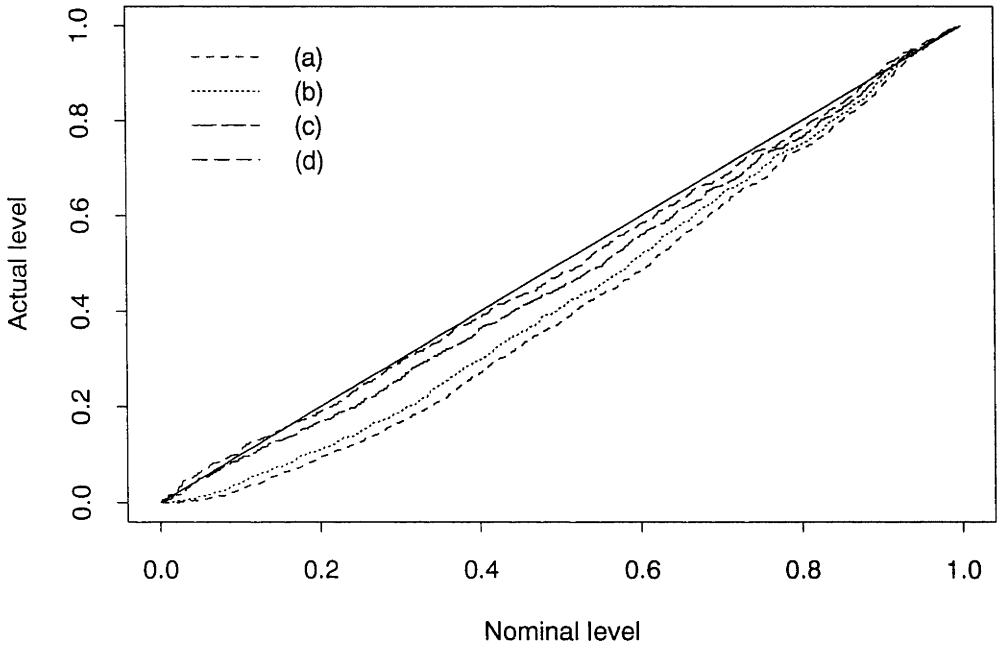


Figure 3.11: *Plot of the level accuracy of forms (a)-(d) of the test for a Normal sample of size* $n = 1000$.

Figure 3.11 contains the same graph as those in Figure 3.10 for a sample size of $n = 1000$. For samples of this size the test has almost attained it asymptotic performance and the asymptotic correction performs very well. In this situation the variance correction makes little improvement to the level accuracy of the test since as the sample size increases the effect of the variance inflation becomes negligible.

3.6 Performance of the corrections

To ascertain whether method (d) provides a test with very good level accuracy for unimodal distributions in general, and not just the Normal, we simulated the test on a wide range of unimodal distributions. The simulations were performed in the same manner as in the last section. The six unimodal densities that we consider are:

(i) the Epanechnikov density;

(ii) the Beta(4,4) density;

(iii) the Beta(3,4) density;

(iv) the Beta(2,4) density;

(v) an Extreme Value density; and

(vi) the Gamma(3) density.

Densities (i)-(iv) have compact support and densities (v) and (vi) have unbounded support. Densities (i) and (ii) are symmetric about their mode and densities (iii)-(vi) are skewed. Densities (ii), (iii), (v) and (vi) have two points of inflexion, density (iv) has one and density (i) has none. The results of these simulations are presented in Figures 3.12–3.17 in the same format as for the Normal distribution (see Figure 3.10).

The first results that we examine are those for the Epanechnikov density. This density is plotted in Figure 3.12 (a) and the other three panels contain plots of the test's level accuracy analogous to those in Figure 3.10. The Epanechnikov density is

$$f(x) = \begin{cases} (3/4)\,(1-x^2) & \text{if } -1 \leq x \leq 1 \\ 0 & \text{otherwise.} \end{cases}$$

The test is far less conservative for the Epanechnikov density than for the Normal. For this reason method (d) tends to overcorrect the conservatism of the test and produces a test where the actual level exceeds the nominal level. The good performance of method (c) for all three sample sizes suggests that the test is performing close to its asymptotic behaviour. For the Normal distribution we saw that even for a sample size of $n = 1000$ the test had not quite reached its asymptotic behaviour. While method (d) does not produce a test with as accurate level as it does for the Normal, and all the densities that follow in this study, the level error for method (d) is by no means disastrous. Cheng and Hall (1997) note that when the excess mass test is performed on the Epanechnikov density, the excess mass test is also less conservative than when it is performed on other densities.

Figures 3.13–3.15 contain the results obtained from performing the test on samples generated from the Beta(4,4), Beta(3,4) and Beta(2,4) distributions respectively. The densities of all three of these distributions are supported on $[0, 1]$. The

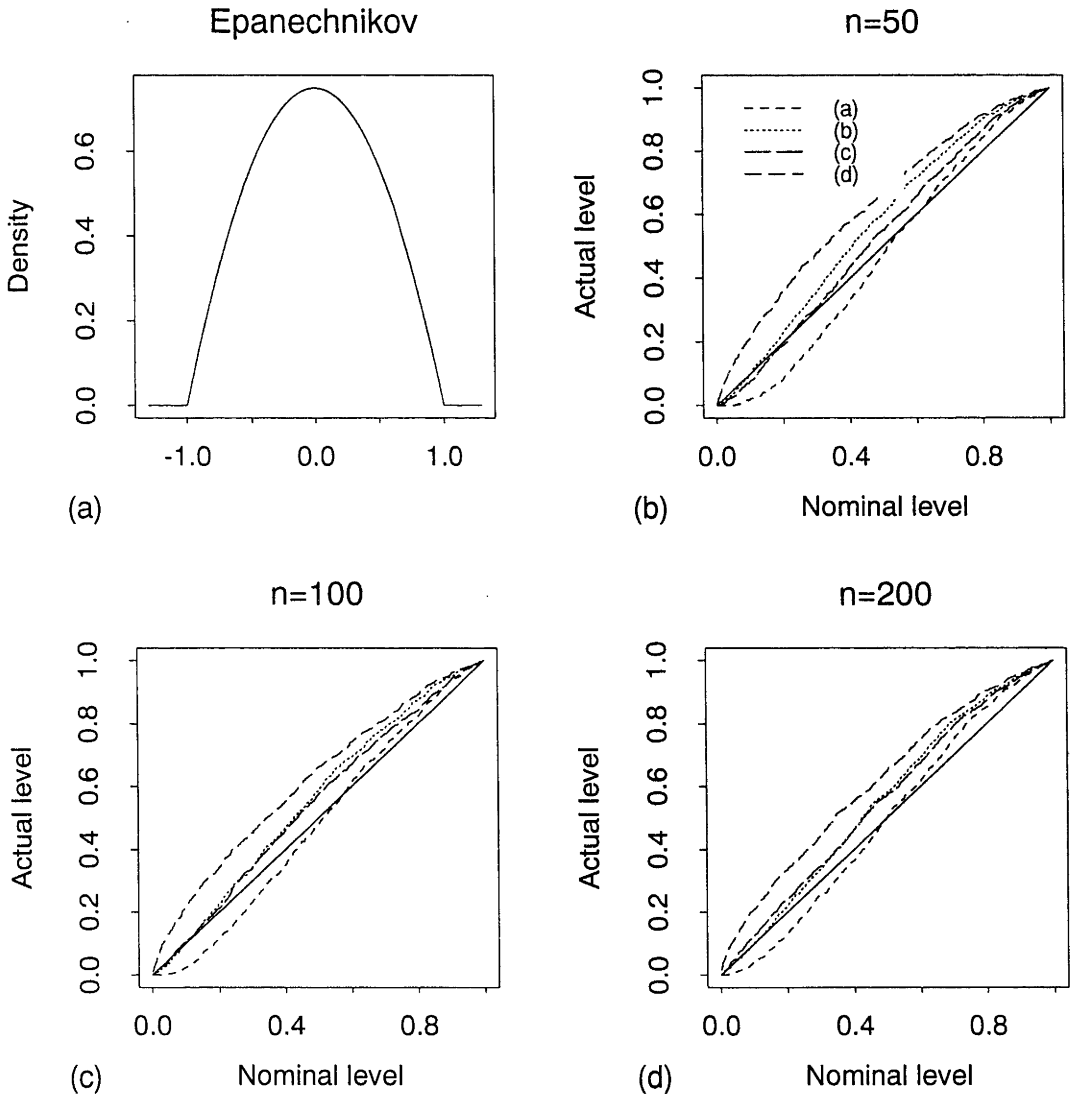


Figure 3.12: *Plot of the level accuracy of forms (a)–(d) of the test. Panel (a) is the sampling distribution, the Epanechnikov. Panels (b)–(d) give the level accuracy for samples of size* $n = 50$, $n = 100$ *and* $n = 200$.

Beta(4,4) density is symmetric, the Beta(3,4) is slightly skewed and the Beta(2,4) is more heavily skewed.

Although the support of the Beta distribution is bounded, tail effects can still be a problem, as observed in Section 2.4. In particular, for a Beta(β_1, β_2) distribution the extreme order statistics are distance n^{-1/β_1} apart in the lower tail and distance n^{-1/β_2} apart in the upper tail. Hence, $\hat{h}_{\text{crit}}$ will be of size $n^{-1/\beta}$, where $\beta = \max(\beta_1, \beta_2)$, if $\beta > 5$. In the examples that we consider $\beta = 4$, which although less than 5, is close enough to cause problems, particularly with small samples.

In these simulations we take $\mathcal{I}$ to be [0,1], the whole of the support of the Beta density. The observations of the previous paragraph indicate that by testing on an appropriate interval $\mathcal{I}$, which is properly contained in (0,1), we might obtain improved performance. We experimented with using different intervals and found that there was little difference in the performance of the test for a range of choices of $\mathcal{I}$. In fact, the performance generally deteriorated a little as the interval was shortened from [0,1].

When the test is performed on the Beta(4,4) distribution the results are quite similar to those for the Normal distribution, which is as one would expect given the similarity of their densities. Once again, method (a) is the most conservative form of the test and method (b) is less conservative than (c) for $n = 50$ but the reverse is true for larger sized samples. The difference from the results of the Normal simulations is that method (d) produced a test with extremely accurate level for the Normal, whereas here method (d) has a tendency to overcorrect the conservatism of the test, producing a test whose actual level is greater than the nominal one; but only slightly so.

The results for the Beta(3,4) distribution are practically identical to those for the Beta(4,4) distribution and require no further comments. The Beta(2,4) results are again quite similar but the level accuracy for method (d) is marginally worse, particularly for $n = 50$. This decreased conservatism is probably due to the test detecting spurious modes in the relatively longer tail. The effects are most marked for $n = 50$ since this is when the data will be most sparse in the tails.

The next distribution that was considered in the simulation study was the Extreme Value distribution with density

$$f(x) = e^{-x} \exp(-e^{-x}), \qquad \text{for } -\infty < x < \infty.$$

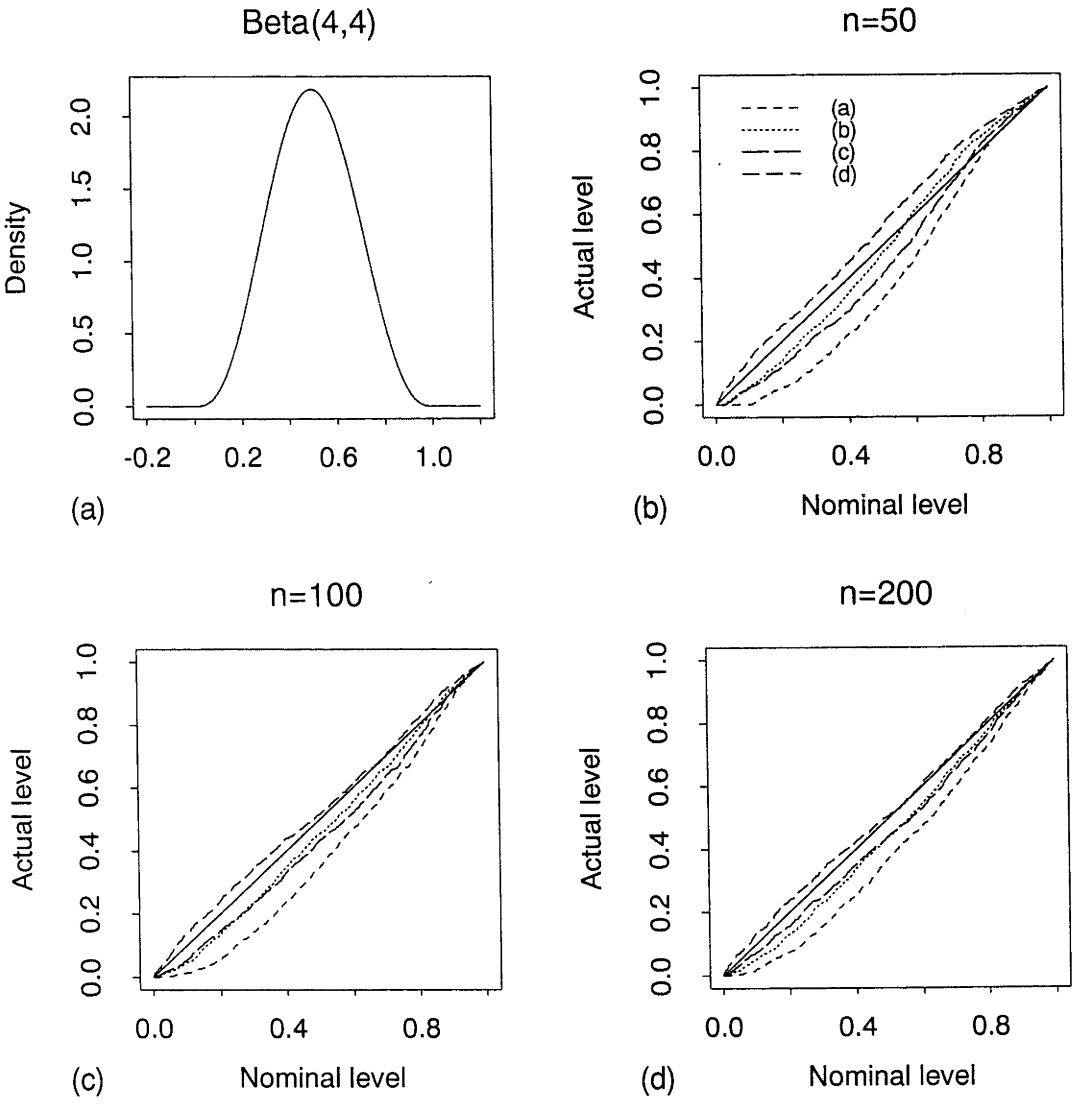


Figure 3.13: *Plot of the level accuracy of forms (a)-(d) of the test. Panel (a) is the sampling distribution, the Beta(4,4). Panels (b)-(d) give the level accuracy for samples of size* $n = 50$, $n = 100$ *and* $n = 200$.

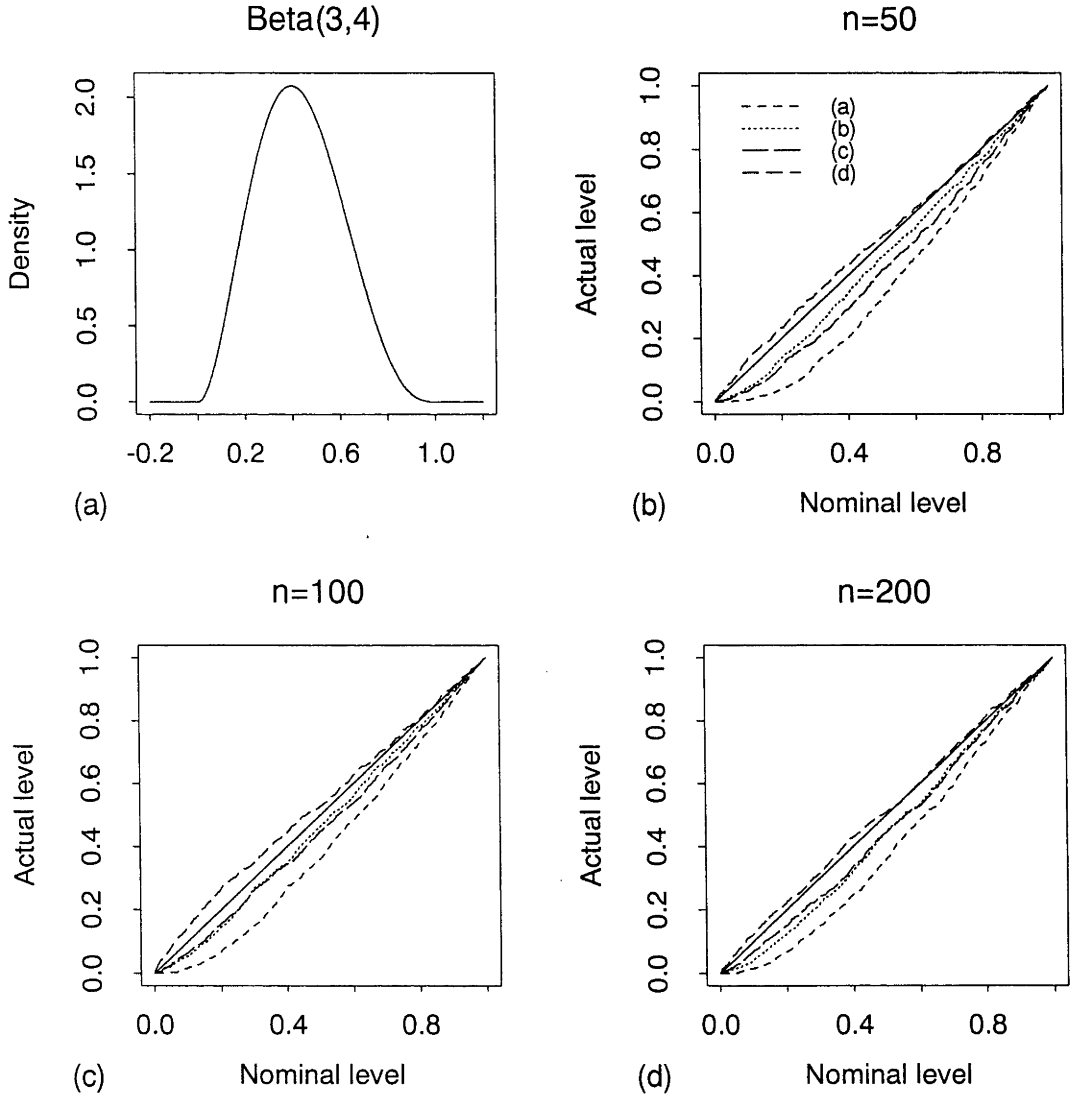


Figure 3.14: *Plot of the level accuracy of forms (a)-(d) of the test. Panel (a) is the sampling distribution, the Beta(3,4). Panels (b)-(d) give the level accuracy for samples of size* $n = 50$, $n = 100$ *and* $n = 200$.

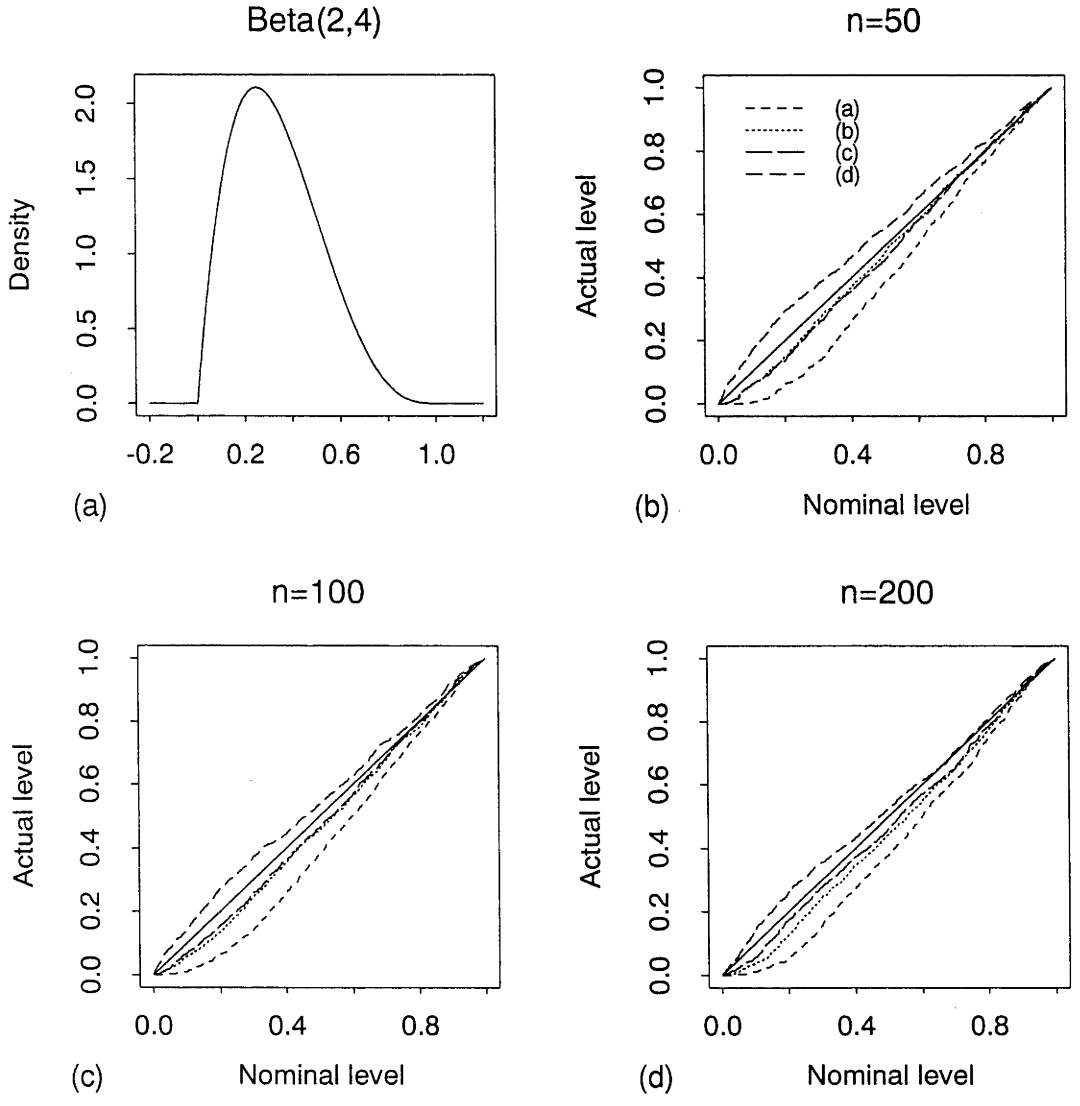


Figure 3.15: *Plot of the level accuracy of forms (a)-(d) of the test. Panel (a) is the sampling distribution, the Beta(2,4). Panels (b)-(d) give the level accuracy for samples of size* $n = 50$, $n = 100$ *and* $n = 200$.

Since this density has infinite support we chose the compact interval $\mathcal{I} = [-1.5, 2.5]$ over which to test the null hypothesis of unimodality. This interval was selected as it contains the bulk of the density's probability mass (just over 0.9) without including the flatter parts of the tails. The test performs very similarly to its Normal distribution behaviour. It can be seen in Figure 3.16 that method (d) produces a test with excellent level accuracy and that the other three forms of the test perform as they usually do.

The final distribution that was investigated was the Gamma(3) distribution. Its density is plotted in Figure 3.17 (a). Like the previous density it has unbounded support and is skewed. In this case, the interval $\mathcal{I}$ was taken to be $[0.5, 5]$. Once the again method (d) produced results with very good level accuracy, for the important levels $\alpha < 0.25$. However the results were not quite as good as they were for the Extreme Value distribution, particularly for the smallest sample size of $n = 50$.

In our simulation study we have observed the performance of four different forms of the critical bandwidth test on the Normal distribution and six other unimodal distributions. For all these distributions, form (a) was the most conservative and (d) was the least. Forms (b) and (c) tended to perform quite similarly for sample size $n = 50$ and form (b) tended to become more conservative relative to form (c) as the sample size increased. Form (d) produced a test with good level accuracy for most of the distributions. The exceptions were the Epanechnikov density and the Beta(2,4), where method (d) tended to overcorrect for the conservatism of the test and produced a test that was inclined to reject the null hypothesis of unimodality with a probability that exceeded the nominal level.

3.7 Calibrating the test for finite sample sizes

We will examine one more technique for correcting the conservatism of the test in this section. In Chapter 2 we discussed two ways of calibrating Silverman's test to improve its level accuracy. The first involved identifying the limiting distribution of the test statistic and adjusting the test so as to give it asymptotically correct level. This is the asymptotic calibration technique – method (d) – whose performance we studied in the previous section. The second method employs Monte Carlo methods based on simulation from a unimodal density. This involves drawing samples of size n from a distribution whose density is unimodal, and applying the bootstrap test to

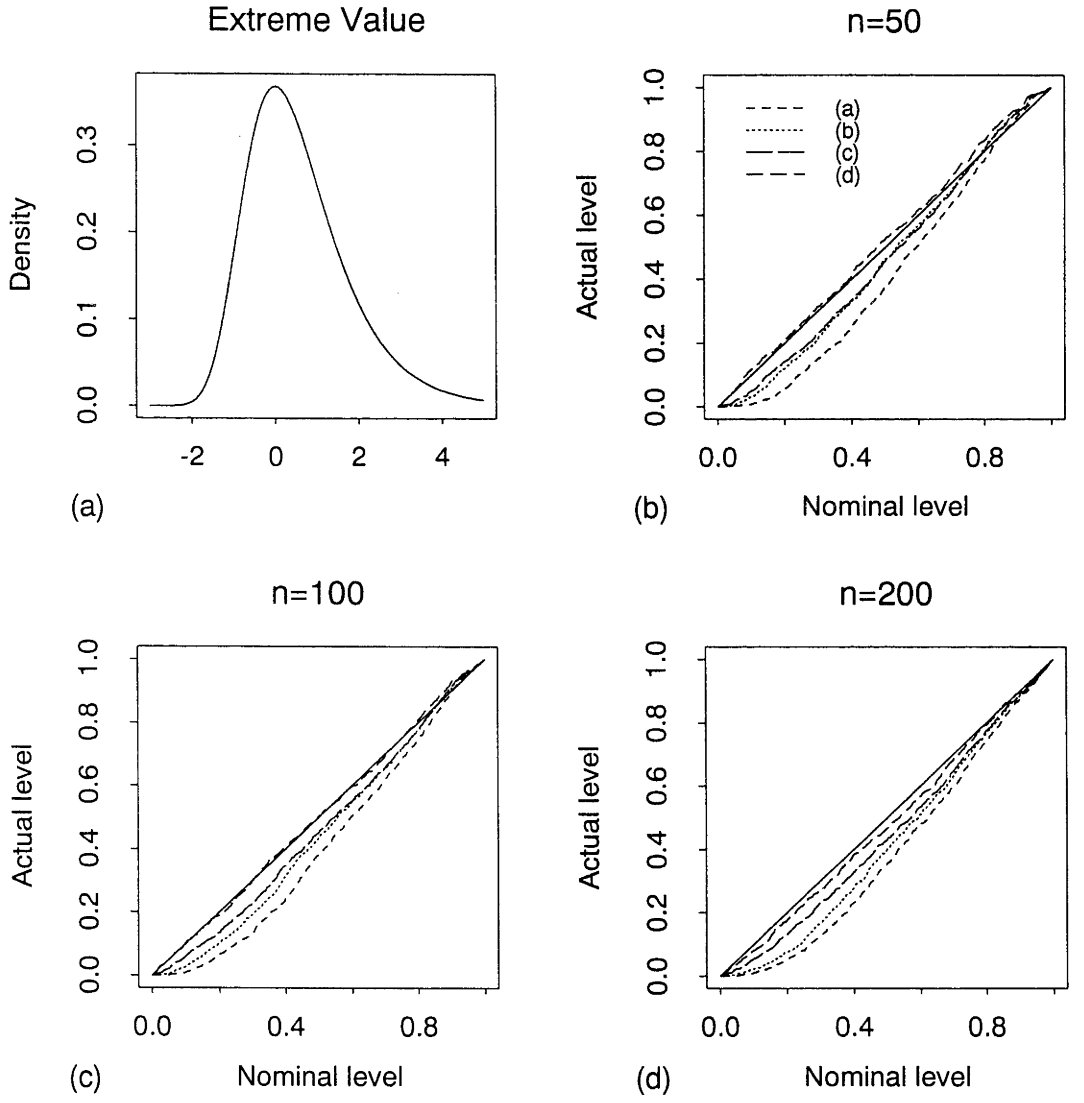


Figure 3.16: *Plot of the level accuracy of forms (a)–(d) of the test. Panel (a) is the sampling distribution, an Extreme Value. Panels (b)–(d) give the level accuracy for samples of size* $n = 50$, $n = 100$ *and* $n = 200$.

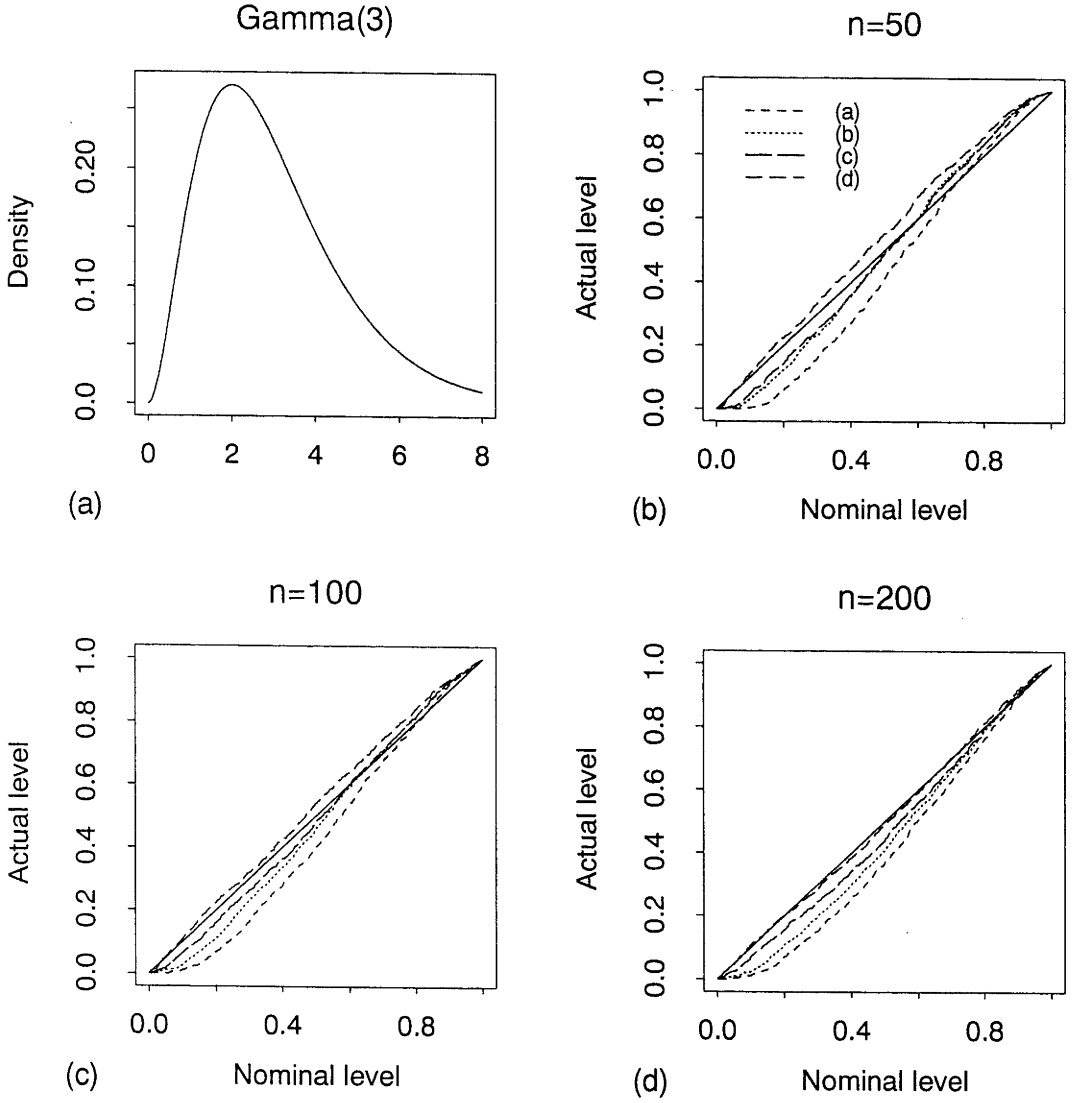


Figure 3.17: *Plot of the level accuracy of forms (a)-(d) of the test. Panel (a) is the sampling distribution, the Gamma(3). Panels (b)-(d) give the level accuracy for samples of size* $n = 50$, $n = 100$ *and* $n = 200$.

each sample. Refer to Section 2.3 for a discussion of this technique. We will refer to this technique as form (e) of the test.

The first way is based purely on asymptotic arguments and is applicable to all distributions. The approach, based on Monte Carlo techniques, is potentially better able to correct for second order effects by taking into account the influence of a particular sample size. However, for finite samples, the value of the adjustment λ_α will depend on which unimodal distribution is used in the simulations to calculate λ_α. Therefore it is necessary to chose a calibration distribution that will provide a value of λ_α that produces a test with good level accuracy when the test is conducted on a wide range of unimodal distributions.

One candidate for the calibration distribution is the Normal distribution. We have seen in the previous section that the test behaves very similarly whether it is performed on the Normal distribution or most of the other unimodal distributions that we have considered. However the test was most conservative when it was applied to samples from a Normal distribution. This means that if the Normal distribution is used to calculate λ_α then, when the test is conducted on data from non-Normal distributions, the calculated value of λ_α will tend to overcorrect for the conservatism of the test. Since the test mostly behaves very similarly whether performed on a Normal sample or a non-Normal sample this overcorrection will usually be quite small and we shall have a test with very good level accuracy.

Another possible choice for the calibration density could be the Epanechnikov. Since the test performs at its least conservative when conducted on samples from the Epanechnikov distribution then the value of λ_α, calculated from this density, will always produce a test with improved level accuracy that will usually still be conservative. We investigated this approach but found that it usually performed no better than the purely asymptotic approach – method (c). As we saw in the previous section the test behaves rather atypically when conducted on samples from the Epanechnikov density so the value of λ_α, based on simulated data from this density, usually does not produce a test with good level accuracy for other distributions.

For these reasons we chose the Normal density to calibrate the test. The result is a test that usually has good level accuracy but will occasionally overcorrect for the conservatism. If a test with reasonable level accuracy but which is still conservative is required then it is best to use the asymptotic calibration. In the remainder of this section we shall give the values of λ_α, calculated using the Normal distribution, for

Table 3.5: *The coefficients for the approximating equation* (3.3) *for a range of sample sizes.*

	$n = 20$	$n = 50$	$n = 100$	$n = 200$	$n = 500$	$n = 1000$	$n = 2000$
a_1	0.73339	0.91662	0.96532	1.00684	0.94766	0.98813	0.95709
a_2	0.23590	-0.46463	-0.40836	-0.85508	-0.39171	-0.84613	-0.47675
a_3	-1.01424	-0.51542	-0.69107	-0.27497	-0.57156	-0.33177	-0.46285
a_4	0.02119	-0.00601	-0.01560	-0.00050	-0.00583	-0.00051	-0.02246
a_5	-0.26157	-0.65929	-0.55643	-0.89999	-0.52350	-0.90848	-0.57447
a_6	0.80542	-0.41150	-0.58990	-0.22520	-0.49397	-0.28207	-0.41155
a_7	-0.01398	-0.00431	-0.01165	-0.00039	-0.00467	-0.00041	-0.01906

a range of sample sizes and we shall investigate the performance of this calibration.

In Section 3.2 we explained how the asymptotic version of λ_α was calculated. There we approximated the asymptotic value of λ_α by simulating the test on samples of size $n = 10000$ drawn from a Normal distribution. In the course of those calculations we calculated λ_α for a range of smaller sample sizes to show that λ_α had converged to its asymptotic value by the time n had reached 10000. For each of these sample sizes we fitted a function of the form

$$\lambda_\alpha = \frac{a_1\alpha^3 + a_2\alpha^2 + a_3\alpha + a_4}{\alpha^3 + a_5\alpha^2 + a_6\alpha + a_7}, \tag{3.3}$$

to provide a means of approximating λ_α for arbitrary α. These functions were plotted in Figure 3.2. The fitted coefficients $a_1, a_2, \ldots, a_7$ were given for $n = 10000$ in Table 3.1. In Table 3.5 we give these fitted coefficients for $n = 20$, 50, 100, 200, 500, 1000 and 2000.

To assess the performance of this sample-size based calibration we trialled these values of λ_α on the same distributions that we studied in the previous section. The results are presented in the same format as before. On the plots of level accuracy we have shown the levels of the test for the sample-size based calibration – form (e) of the test – along with the results for the asymptotic calibration – form (c) – and the asymptotic calibration with the resampled data rescaled to have the same variance as the original data – form (d).

The first distribution we consider is the Normal and the simulation results are plotted in Figure 3.18. Form (e) of the test performs extremely well, which is

exactly what one would expect since we used the Normal distribution to calculate λ_α. The results for the six other distributions considered in this study are plotted in Figures 3.19–3.24. For the distributions for which method (d) performed extremely well, namely the Extreme Value and the Gamma(3), method (e) performed slightly worse than method (d). Otherwise method (e) performs better than (d) particularly for the distributions on which (d) did not perform so well, the Epanechnikov and the Beta(2,4). On the whole method (e) produced a test with very good level accuracy. It was the best method for the Normal and the three Beta distributions and was very close to being the best for the Extreme Value and the Gamma(3), where it was just outperformed by form (d). For the Epanechnikov density method (c) produced the most accurate level but form (e) was substantially better than form (d).

We will complete this section with some recommendations about the best form of the test to use in different situations. In general, (e) is the best form of the test to use. The value of λ_α is easily calculated by using the coefficients given in Table 3.5. For sample sizes not given in the table the value of λ_α can be interpolated, or it is simpler, and probably better in terms of level accuracy, to use the value for the immediately larger sample size that is given in the table. The use of λ_α for a slightly larger sample size will have little effect on level accuracy and what effect it has will be in the interests of conservatism.

While method (e) generally produces the most accurate level there are times when it will overcorrect for the conservatism of the test; see Figure 3.19. If a conservative test is required then form (c) is the method that produces a test with the best level accuracy, amongst those forms of the test that are consistently still slightly conservative. The only exception is that for sample sizes smaller than $n = 50$ our simulations reveal that form (b) outperforms form (c).

3.8 Power of the test

We complete this chapter by examining the power of the various forms of the test. The purpose of calibrating a conservative test to give it correct level accuracy is to increase its power, particularly in difficult testing situations where the modes are not well separated. We examined the performance of the test on a wide range of bimodal distributions, most of which were mixtures of Normals, as well as some mixtures of Beta distributions. In every test in this section we are testing the null

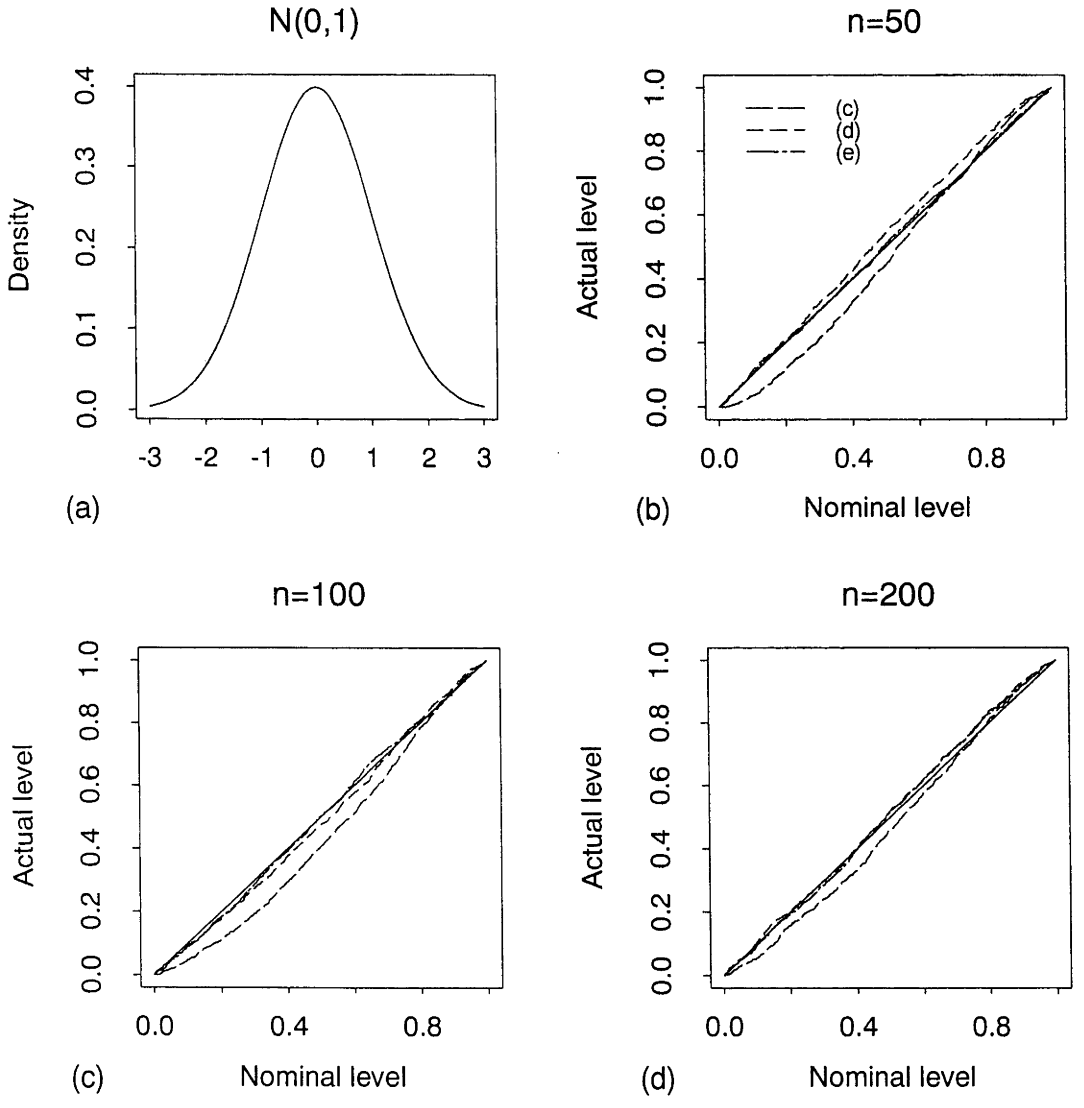


Figure 3.18: *Plot of the level accuracy of forms (c)-(e) of the test. Panel (a) is the sampling distribution, the standard Normal. Panels (b)-(d) give the level accuracy for samples of size* $n = 50$, $n = 100$ *and* $n = 200$.

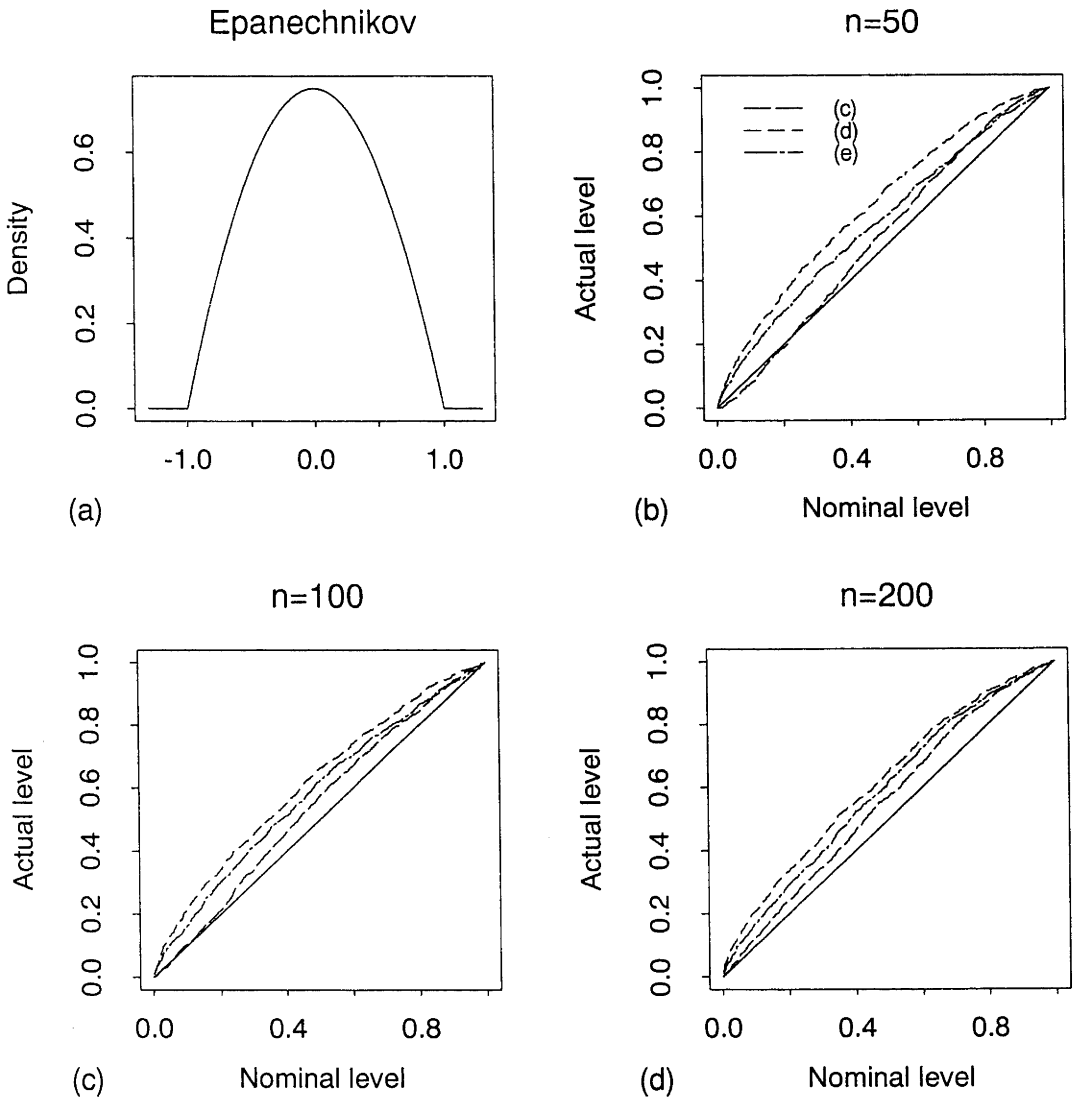


Figure 3.19: *Plot of the level accuracy of forms (c)-(e) of the test. Panel (a) is the sampling distribution, the Epanechnikov. Panels (b)-(d) give the level accuracy for samples of size* $n = 50$, $n = 100$ *and* $n = 200$.

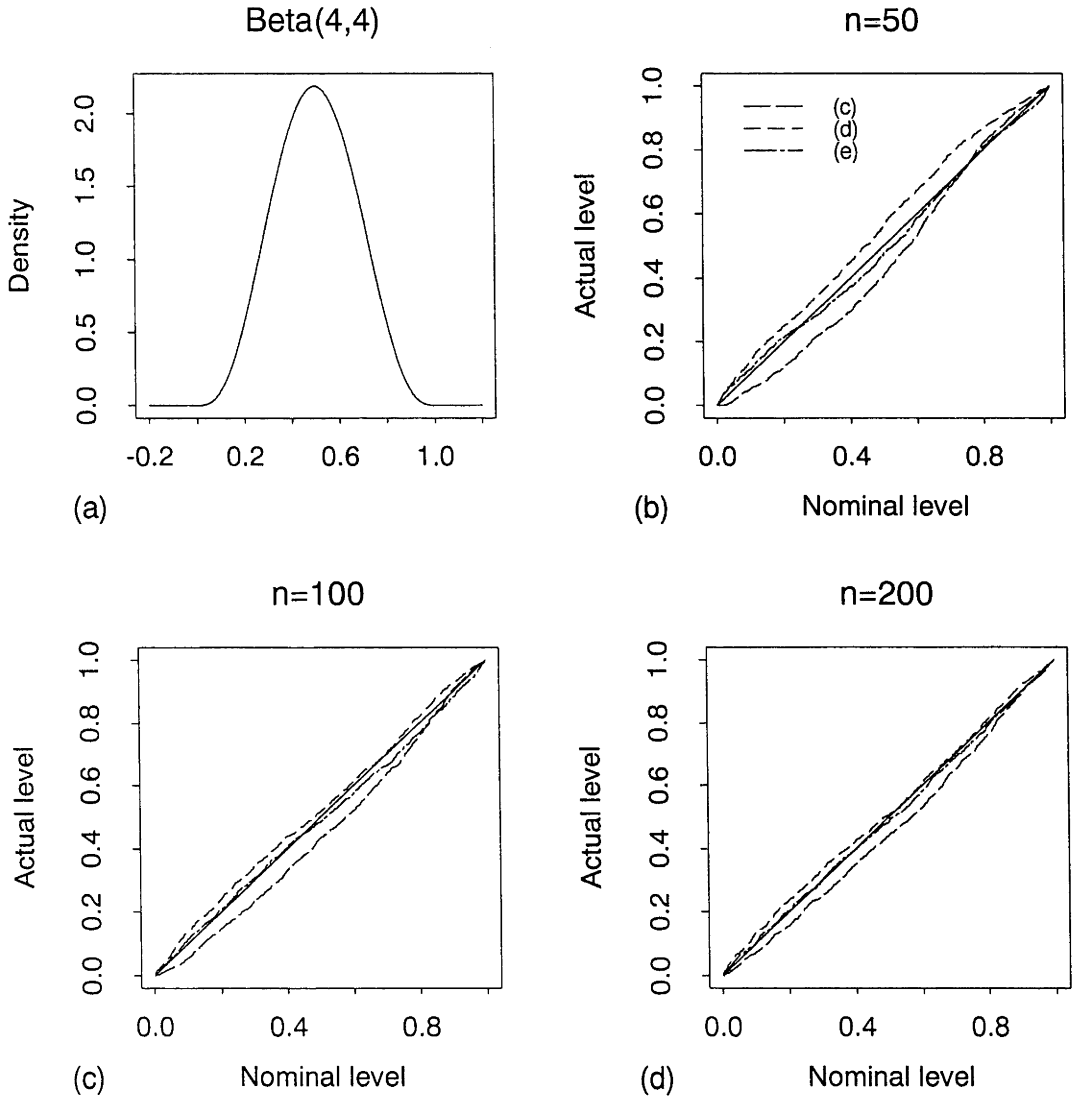


Figure 3.20: *Plot of the level accuracy of forms (c)-(e) of the test. Panel (a) is the sampling distribution, the Beta(4,4). Panels (b)-(d) give the level accuracy for samples of size* $n = 50$, $n = 100$ *and* $n = 200$.

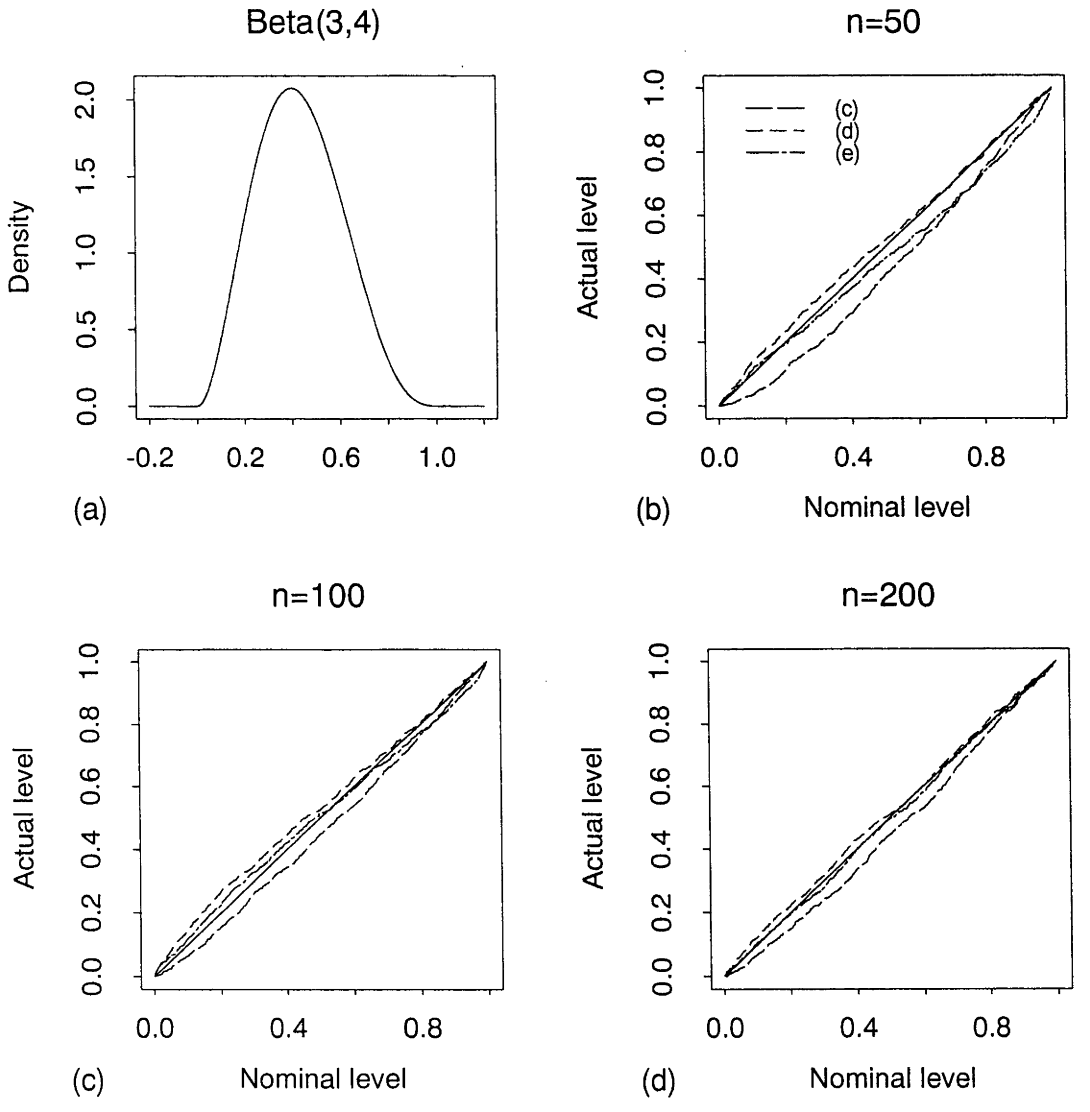


Figure 3.21: *Plot of the level accuracy of forms (c)-(e) of the test. Panel (a) is the sampling distribution, the Beta(3,4). Panels (b)-(d) give the level accuracy for samples of size* $n = 50$, $n = 100$ *and* $n = 200$.

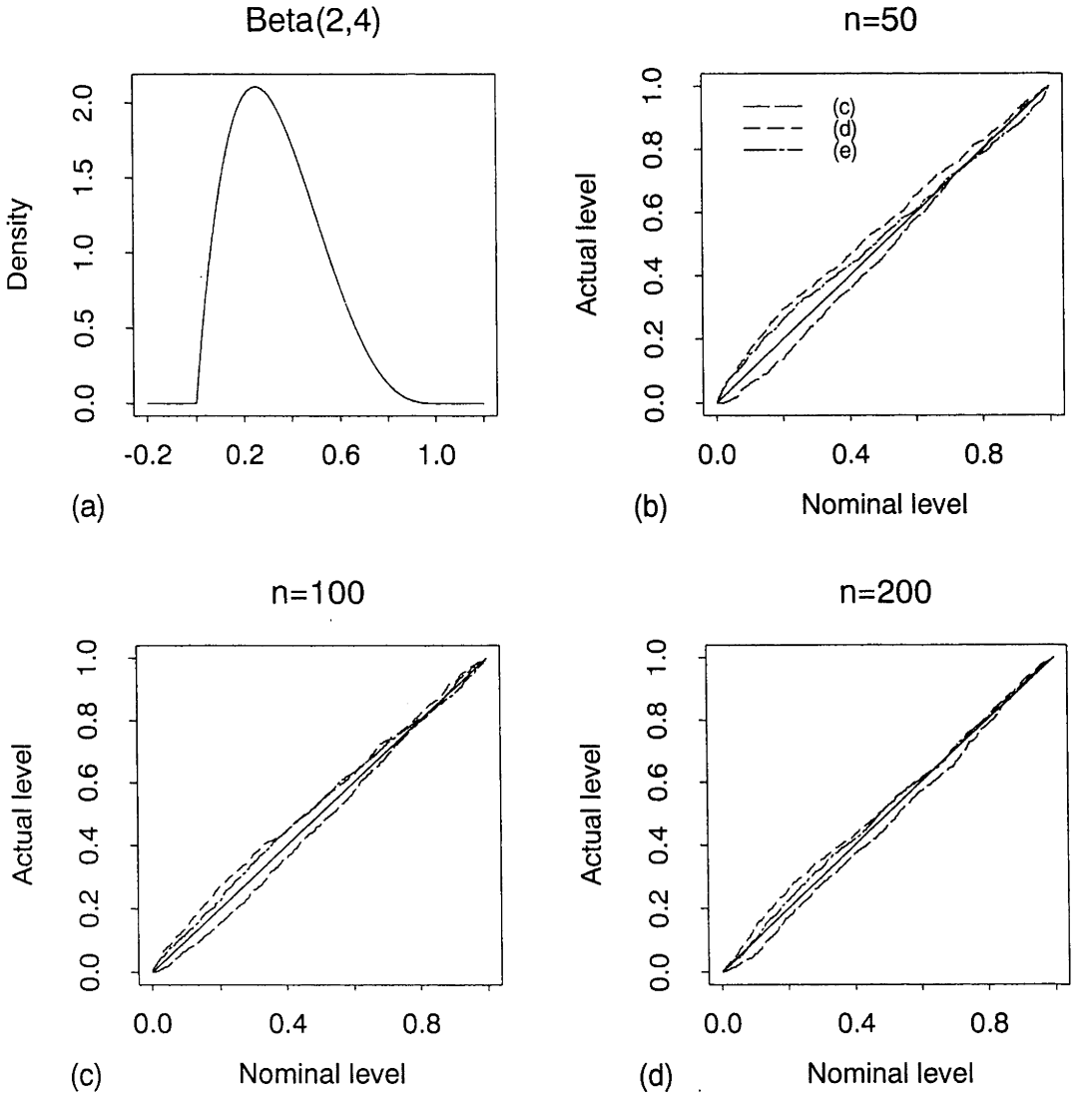


Figure 3.22: *Plot of the level accuracy of forms (c)-(e) of the test. Panel (a) is the sampling distribution, the Beta(2,4). Panels (b)-(d) give the level accuracy for samples of size* $n = 50$, $n = 100$ *and* $n = 200$.

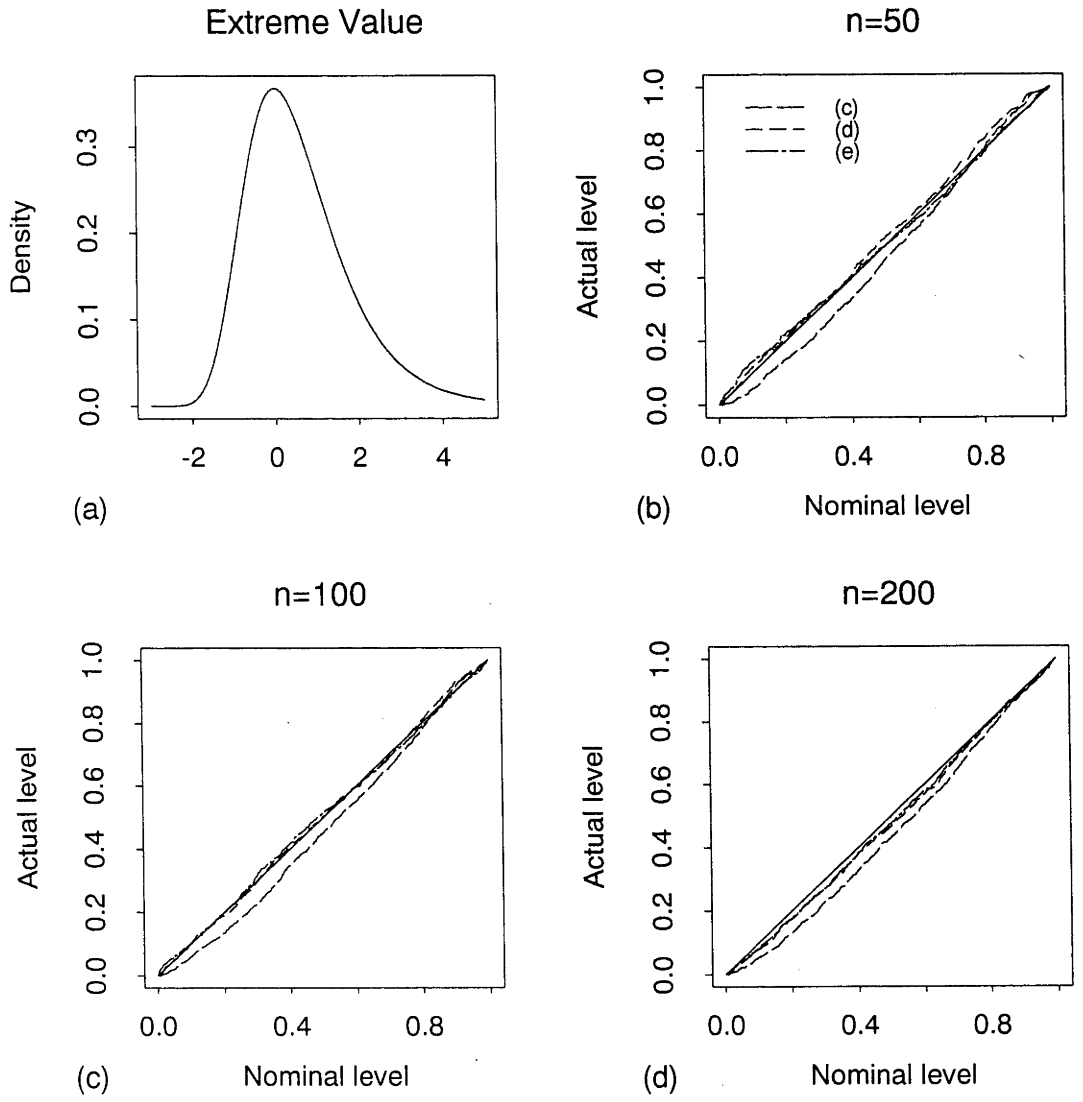


Figure 3.23: *Plot of the level accuracy of forms (c)–(e) of the test. Panel (a) is the sampling distribution, an Extreme Value. Panels (b)–(d) give the level accuracy for samples of size* $n = 50$, $n = 100$ *and* $n = 200$.

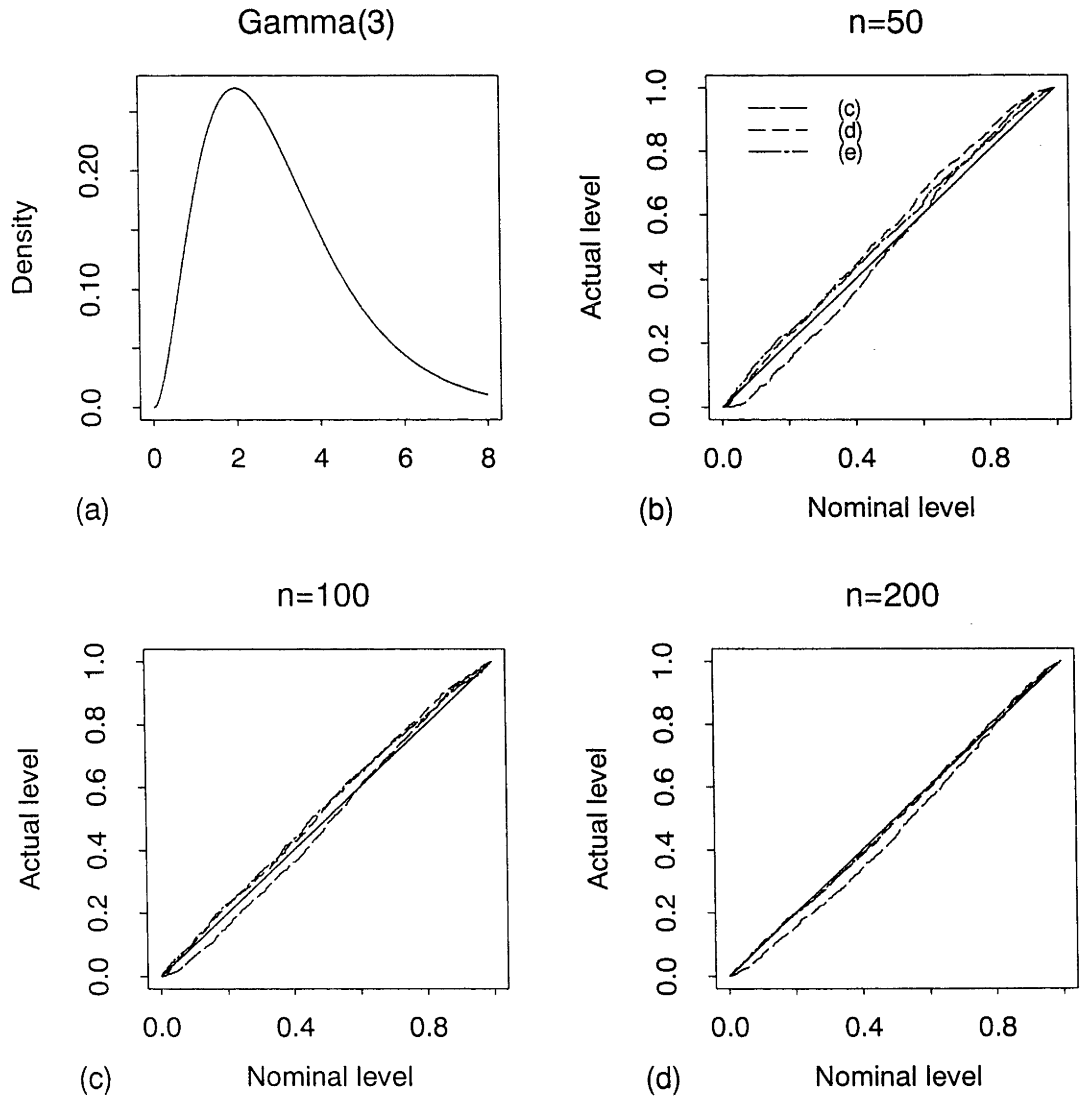


Figure 3.24: *Plot of the level accuracy of forms (c)-(e) of the test. Panel (a) is the sampling distribution, the Gamma(3). Panels (b)-(d) give the level accuracy for samples of size* $n = 50$, $n = 100$ *and* $n = 200$.

hypothesis that the density has precisely one mode against the alternative that it has more than one.

The first group of densities that we consider are the symmetric mixtures of two Normals that have densities of the form

$$f(x) = 0.5\,\phi(x; -\mu, 1) + 0.5\,\phi(x; \mu, 1), \tag{3.4}$$

where $\phi(x; \mu, \sigma^2)$ is the Normal density with mean μ and variance σ^2. Figure 3.25 displays theses densities for the values of $\mu = 1.0, 1.1, 1.2, \ldots, 1.9, 2.0$. For $\mu = 1.0$ the density is unimodal but it is almost bimodal. We have included it in our study since it will give us an indication of the level accuracy of the test in an unusual and difficult situation. For $\mu = 1.1$ the density is slightly bimodal and as μ increases to 2.0 the separation of the modes increases until the density is fairly strongly bimodal.

We carried out forms (a) through (e) of the test for sample sizes $n = 50$ and $n = 200$. Remember from the previous section that form (e) is the most accurate form of the test, followed by (d). Form (b) is the usual version of the test, as proposed by Silverman (1981). Hence we shall mainly concentrate on comparing methods (b) and (e). The other methods are essentially included for completeness. We chose to perform the test over the interval $\mathcal{I} = [-\mu - 1.5, \mu + 1.5]$. When we were performing the test on the Normal distribution we decided it was best to confine the interval $\mathcal{I}$ to within 1.5 standard deviations of the mean, to avoid troubles with data sparsity in the tails of the distribution. In this case we chose the lower limit of the interval to be 1.5 times the standard deviation of the left hand component of the mixture below the mean of that component. The upper limit was similarly chosen for the right-hand component density of the mixture.

In Figure 3.26 the power of the test is plotted against the value of μ for the two sample sizes, $n = 50$ and $n = 200$, for the nominal levels of $\alpha = 0.05$ and $\alpha = 0.1$. Figure 3.26 (a) is the plot for the case of $n = 50$ and $\alpha = 0.05$. As expected, for all five methods the power of the test increases as the separation of the modes increases. The power for method (e) is generally about 15% greater than the power for method (b), even for the well separated modes in the case $\mu = 2.0$. For the difficult cases of $\mu = 1.1, 1.2$ and 1.3, where the modes are not well separated, the calibrated versions of the test, forms (d) and (e), have over twice the power of method (b). The situation is quite similar in Figures 3.26 (b)–(d) but naturally the power increases as α increases from 0.05 to 0.1 and n from 50 to 200.

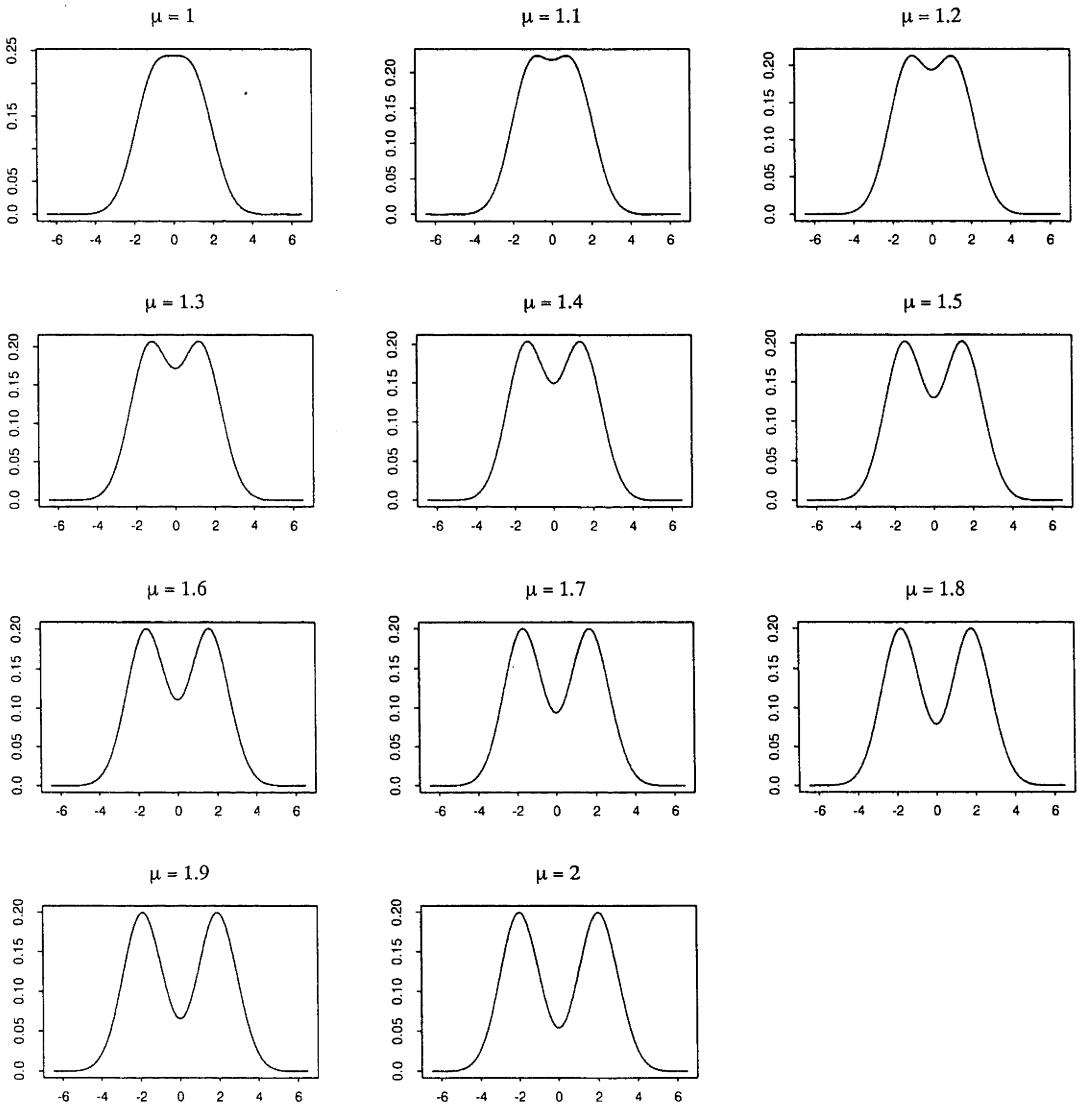


Figure 3.25: *Family of symmetric Normal mixtures with densities of the form* (3.4).

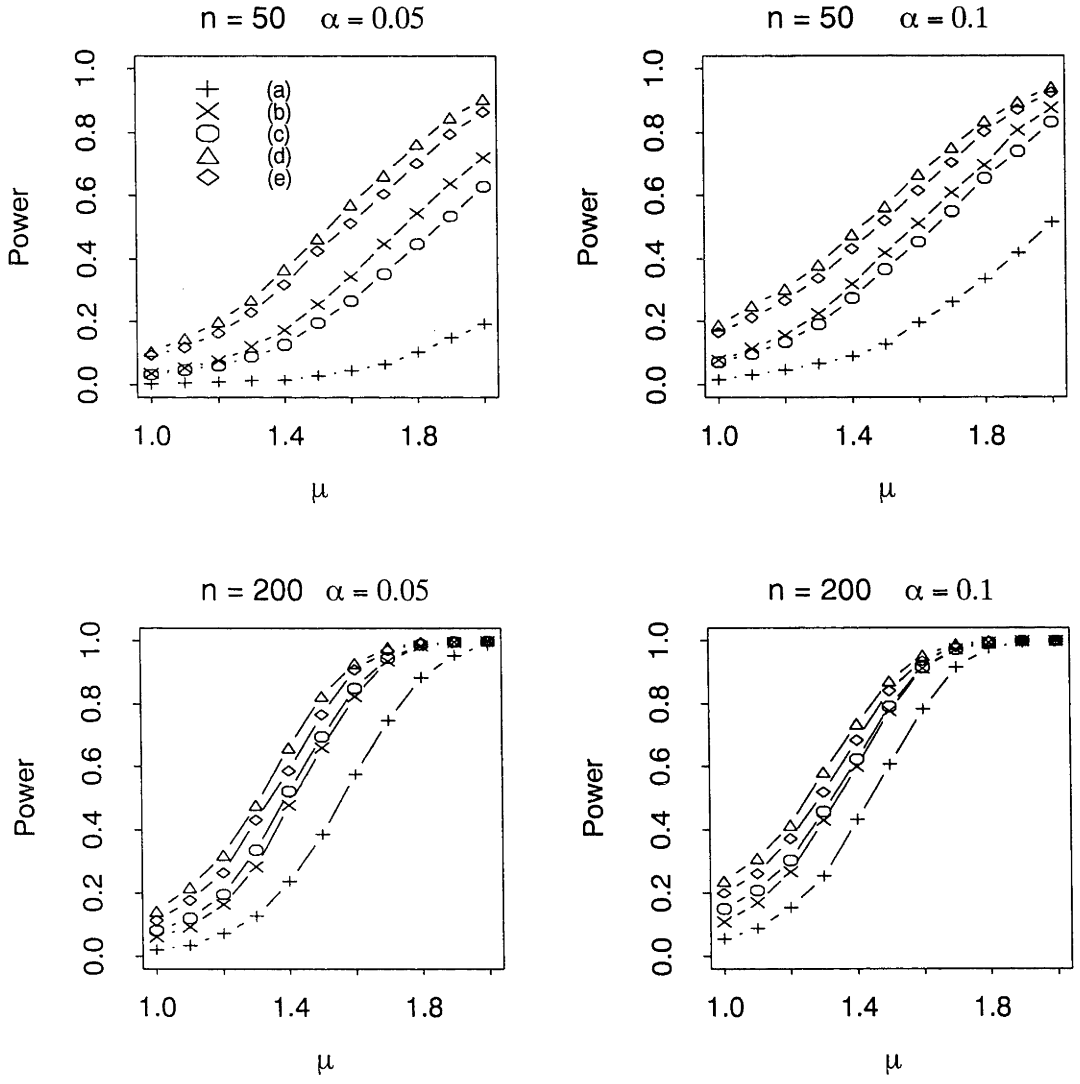


Figure 3.26: *Power of the five forms of the test for the densities* (3.4), *for* $n = 50$ *and* $n = 200$ *and* $\alpha = 0.05$ *and* $\alpha = 0.1$.

For $\mu = 1.0$ the density is actually unimodal, so for $\alpha = 0.05$ we would hope that the probability of rejecting the null hypothesis is close to 0.05. However, our calibration techniques were developed for unimodal densities whose modes are not too flat; in Section 2.6 we imposed the condition that the second derivation of the density at the mode is non-zero. The density for $\mu = 1.0$ has an extremely flat mode; both its second and third derivatives are zero at the mode. This density can be considered to lie on the boundary between the classes of unimodal and bimodal distributions, since for $\mu \leq 1$ the density is unimodal and for $\mu > 1$ it is bimodal. Therefore we would not expect the test level to be entirely accurate for method (e) but we would hope that it is reasonable. For $n = 50$ and $\alpha = 0.05$ the actual level of form (e) is 0.10 and for $\alpha = 0.1$ it is 0.17. For $n = 200$ the actual level is 0.11 for $\alpha = 0.05$ and 0.20 for $\alpha = 0.1$. Considering that this is a worst case scenario, and a rather unlikely one, these results are reasonably tolerable.

The next group of densities in our study are the mixtures of Normals with densities of the form

$$f(x) = \pi\, \phi(x; -3, 1) + (1 - \pi)\, \phi(x; 3, 2^2). \tag{3.5}$$

The densities are displayed in Figure 3.27 for $\pi = 0.1, 0.2, \ldots, 0.8, 0.9$. The bimodality is strongest for about $\pi = 0.4$ and weakest for π near 0 or 1.

As in the previous example we carried out forms (a) through (e) of the test for samples of size $n = 50$ and $n = 200$. We took $\mathcal{I}$ to be the interval [-4.5, 6]. The power of the test is plotted against π in Figure 3.28 for $n = 50$ and $n = 200$, for both $\alpha = 0.05$ and $\alpha = 0.1$. Figure 3.28 (a) gives the power in the case $n = 50$ and $\alpha = 0.05$. The usefulness of our calibrated methods is again evident, particularly in the difficult testing cases. For $\pi = 0.1$ method (e) has over twice the power of method (b), and for other values of π the power of method (e) is normally at least 0.1 greater than the power of method (b).

For $\pi = 0.7, 0.8$ and 0.9 the right-hand mode is not very distinct, yet most forms of the test still deliver reasonable power. Unfortunately this high power is most probably due to the sparsity of the data in the flat right hand mode. However the power of method (e) does steadily decrease as π increases from 0.7 to 0.9, which does suggest that, in this case at least, it deals which the problems of data sparsity better than the other methods. For $n = 200$ these data sparsity problems do not exist and the powers of all forms of the test rapidly decrease as π increases from 0.7 to 0.9.

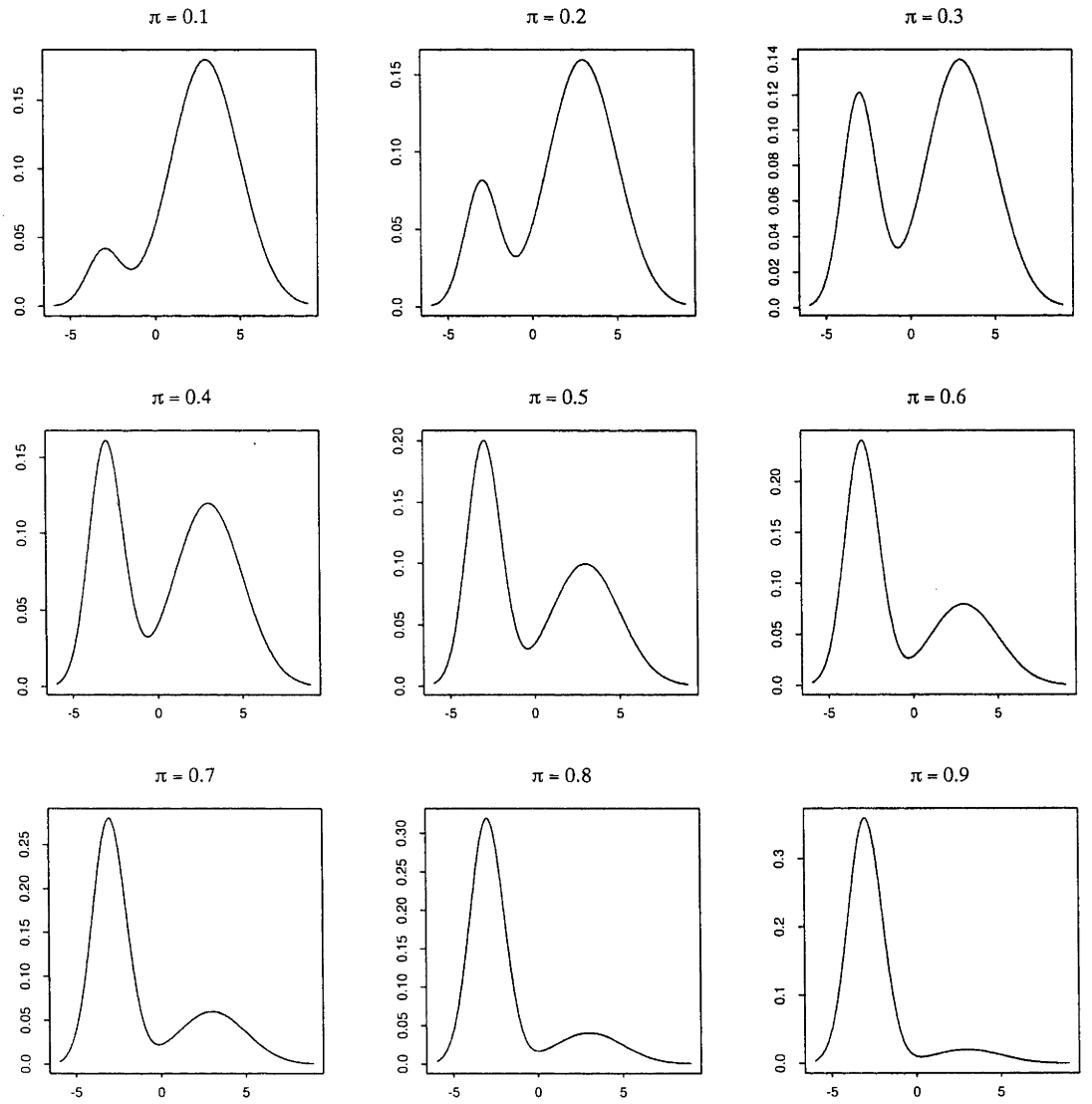


Figure 3.27: *Family of Normal mixture distributions with densities of the form* (3.5).

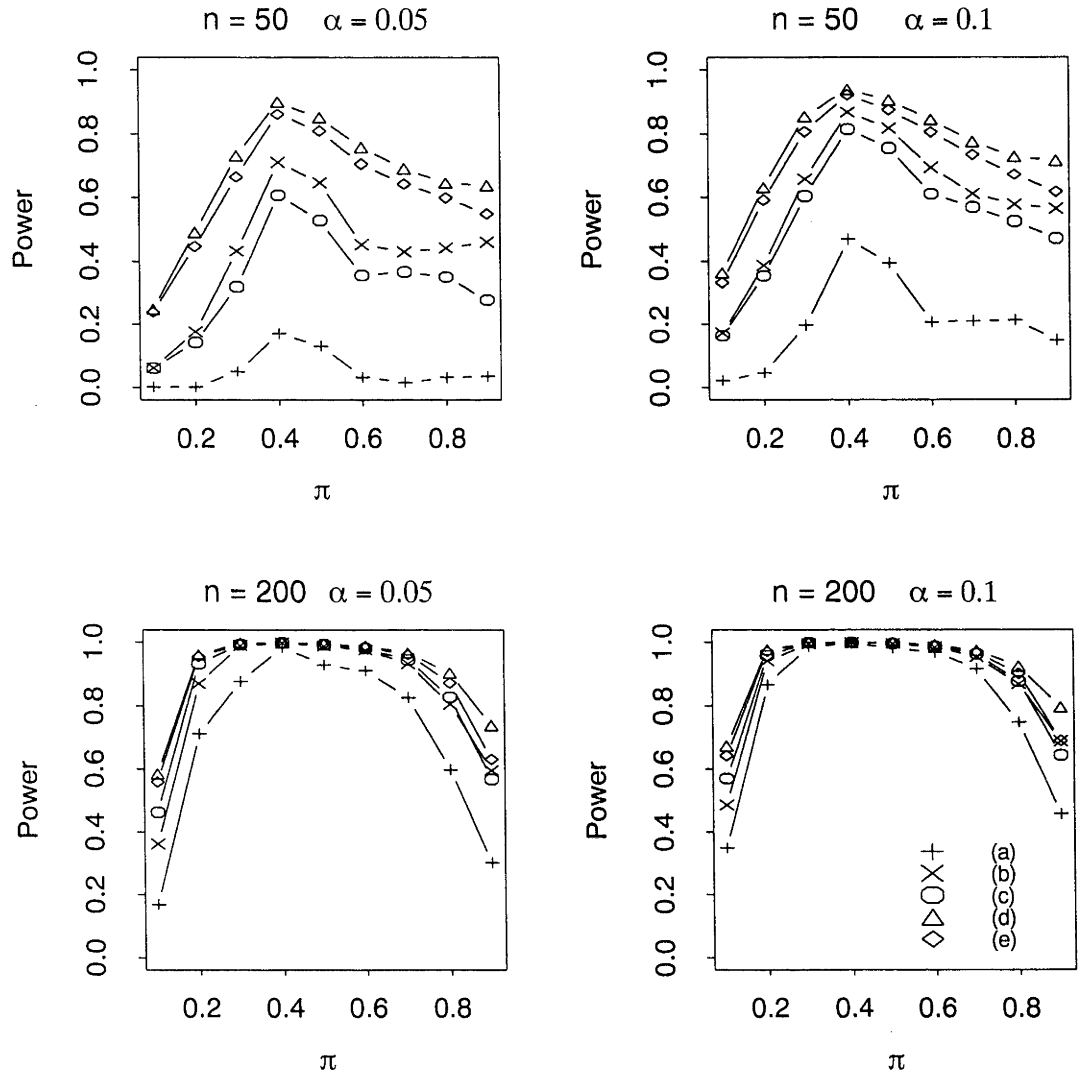


Figure 3.28: *Power of the five forms of the test for the densities* (3.5), *for* $n = 50$ *and* $n = 200$ *and* $\alpha = 0.05$ *and* $\alpha = 0.1$.

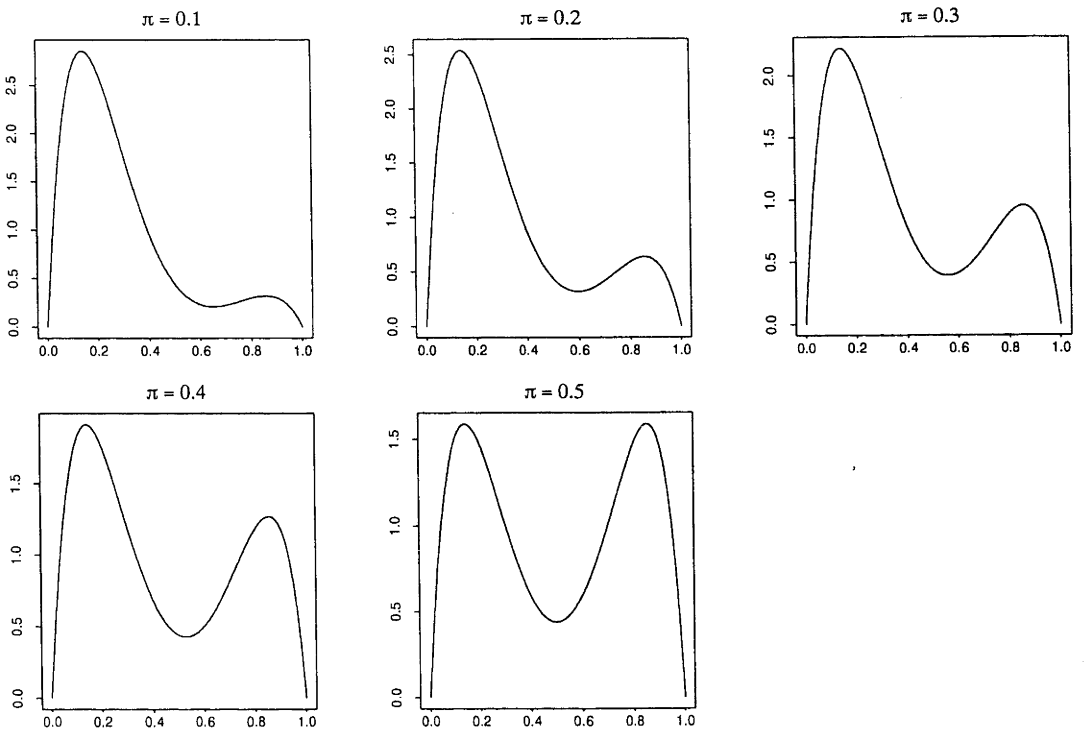


Figure 3.29: *Family of mixtures of a Beta(7,2) and a Beta(2,7) with densities of the form* (3.6).

The final group of mixtures that we consider is that of the following mixtures of Beta distributions:

$$\pi \,\mathrm{Beta}(7,2) + (1-\pi)\,\mathrm{Beta}(2,7). \tag{3.6}$$

These are different from those studied previously in that they have compact support and the component densities are asymmetric. The densities of these mixtures are plotted in Figure 3.29 , for $\pi = 0.1, 0.2, 0.3, 0.4$ and 0.5 .

The power of the test is graphed against π in Figure 3.30 , for $n = 50$ and $n = 200$ for the two levels of $\alpha = 0.05$ and $\alpha = 0.1$. Naturally the power increases with π. For $\alpha = 0.05$ and $n = 50$ the power of method (e) is between 10% and 20% greater than the power of method (b). Unlike the previous examples, calibration is most effective in the cases of $\pi = 0.2$ and $\pi = 0.3$ rather than in the most difficult case of $\pi = 0.1$.

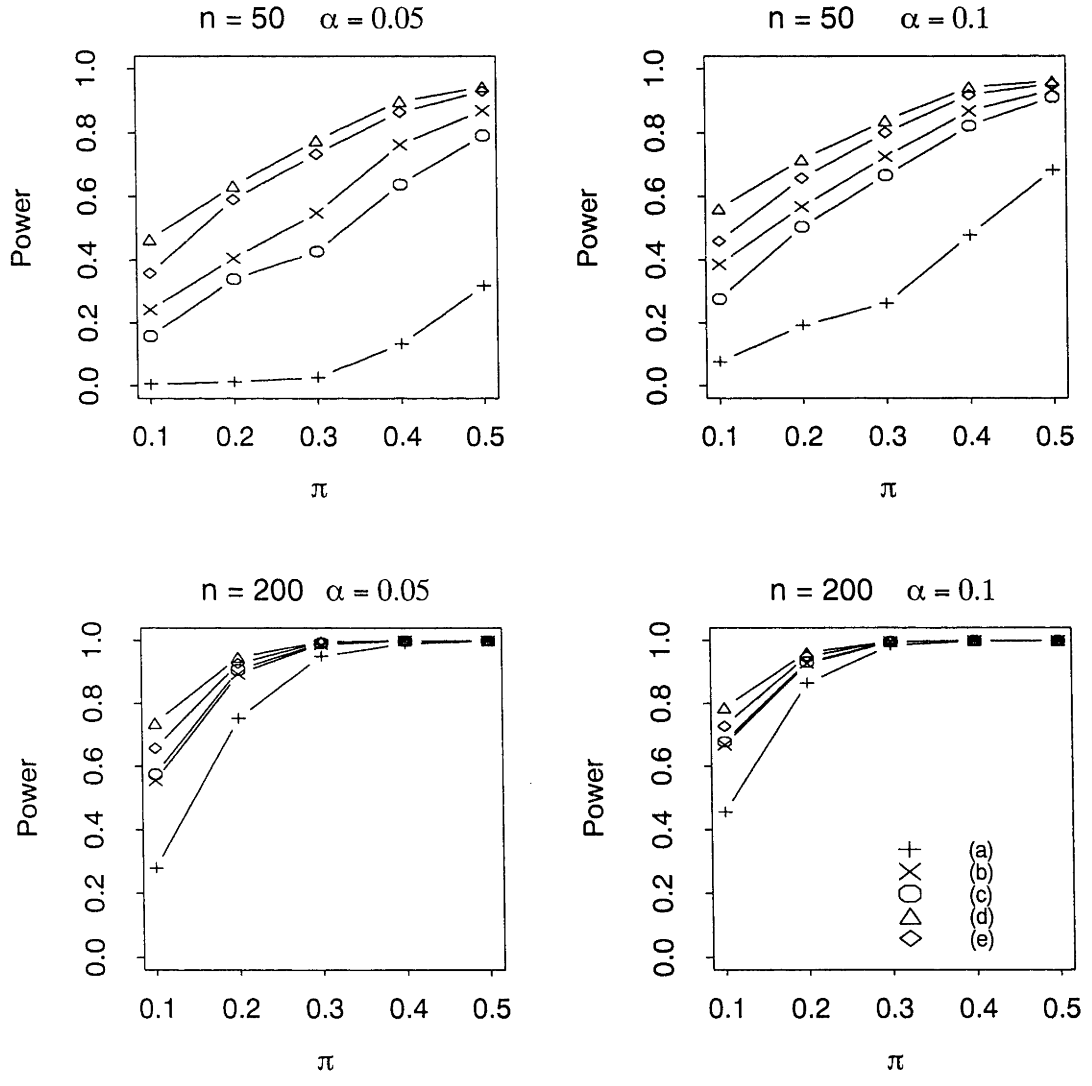


Figure 3.30: *Power of the five forms of the test for the densities* (3.6), *for* $n = 50$ *and* $n = 200$ *and* $\alpha = 0.05$ *and* $\alpha = 0.1$.

Chapter 4

Examples

4.1 Introduction

In this chapter we apply the techniques discussed in Chapters 2 and 3 to real data sets. There were two main modifications to Silverman's test developed in Chapter 2. The first of these was the development of a calibrated version of the test which has greater power and level accuracy than the form of the test proposed by Silverman (1981). The second modification was testing for the number of modes over a compact interval $\mathcal{I}$ instead of over the whole line. This modification was designed to overcome problems caused by spurious modes arising from the sparsity of the data in the tails of a distribution. This method has other uses, which will be seen later in this chapter. The chapter aims to show that these modifications to the critical bandwidth test increase the effectiveness of the test for data analysis.

Throughout this chapter we are interested in testing the null hypothesis

$$H_0 : f \text{ has precisely } j \text{ modes in } \mathcal{I}$$

versus the alternative hypothesis

$$H_1 : f \text{ has } j+1 \text{ or more modes in } \mathcal{I},$$

where $\mathcal{I}$ is some compact interval in which the density f does not vanish. When we make shorthand statements like "we tested for $j = 2$ modes", we mean, more precisely, that we tested the null hypothesis that f has precisely two modes in $\mathcal{I}$ versus the alternative hypothesis that f has three or more modes in $\mathcal{I}$.

In Chapter 3 we studied five different forms of the critical bandwidth test. These were:

(a) not variance corrected and not calibrated;

(b) variance corrected but not calibrated;

(c) not variance corrected but asymptotically calibrated;

(d) variance corrected and asymptotically calibrated; and

(e) not variance corrected but calibrated for a sample size of n.

The simulations conducted in Sections 3.5–3.8 demonstrated that versions (d) and (e) of the test have the greatest power and level accuracy. Form (d) usually has a little more power but (e) tends to have better level accuracy. Forms (b) and (c) are quite conservative. Form (b) of the test is better than form (c) for samples with fewer than 50 observations and form (c) is better for samples of size 100 or greater. The conservatism of form (c) of the test decreases with sample size; asymptotically it produces a test with correct level. Version (a) of the test is always inferior to version (b) and is included only for completeness.

It was explained in Section 2.5 that the calibrated versions of the test, (c), (d) and (e), are only appropriate when the null hypothesis of $j = 1$ mode is being tested. When the null hypothesis of j modes, for $j \geq 2$, is being tested we use the very conservative method (b). Form (b) is the version of the test originally proposed by Silverman. In the analyses presented in the remaining sections of this chapter, we give the p-values for all five versions of the test when testing the null hypothesis of unimodality. We include the results for all five methods to illustrate the differences between the methods but we only rely on the p-values from versions (d) and (e) of the test to decide whether to accept or reject the null.

We tend to reject the null hypothesis of unimodality if the p-value is less than 0.1, but we take a fairly flexible approach. When we are testing for two or more modes we use the very conservative method (b). In this situation we bear this conservatism in mind and lean towards rejecting the null hypothesis if the p-values is less than 0.2.

The interval $\mathcal{I}$ in which we are assessing the modality of the density f is usually chosen to include all the data points. An exception to this is if a data set contains

obvious outliers, when we normally choose $\mathcal{I}$ to exclude these points so that they do not have adverse effects on the testing procedure. These possible effects were discussed in Section 2.4.

While we were motivated by tail problems to study the testing problem over a compact interval rather than over the whole line, there are other advantages to this approach. The effectiveness of the critical bandwidth test can be hampered by its reliance on a kernel density estimator with a single, global bandwidth. For example, densities with modes located on peaks of very different shapes and sizes are not well estimated by a kernel estimator with a single bandwidth, and this can lead to the test producing misleading results. In Sections 4.5 and 4.6 we encounter data sets that display this characteristic. We can overcome this problem by testing for the number of modes over different intervals. This allows the use of different bandwidths for different intervals, thus lending the kernel density estimator a greater degree of local adaptivity. Testing over different intervals can also be useful to clarify results when testing over the bulk of the data does not produce a conclusive result.

If we test over the whole line, our powerful, calibrated versions of the critical bandwidth test are only applicable when testing for unimodality. By testing for unimodality in appropriately chosen intervals we can combine the results from a number of calibrated tests to draw conclusions about the total number of modes.

4.2 Chondrite data

The first data set that we examine consists of measurements of the percentage of silica in 22 chondrite meteors. These data have been considered by many authors. In the context of mode testing they have been analysed by Good and Gaskins (1980), Silverman (1981), Müller and Sawitzki (1991), Minnotte and Scott (1993), Minnotte (1997) and Cheng and Hall (1998). It is universally agreed that these data come from a trimodal distribution.

We chose $\mathcal{I}$ to be the interval [18,37] and we began by considering the case of testing for $j = 1$ mode. For method (e) the size of the calibration depends on the sample size. Instead of recalculating the calibration for a sample size of $n = 22$ we used the values for $n = 20$ that we had previously computed in Chapter 3. The value of the critical bandwidth $\hat{h}_{\text{crit}}$ and the p-values for the five different forms of the test are listed in Table 4.1.

Table 4.1: *Critical bandwidth and p-values for testing the null hypothesis of unimodality on the chondrite data. Versions (a)–(e) of the test were performed on the interval* $\mathcal{I} = [18, 37]$.

$\hat{h}_{\text{crit}}$	(a)	(b)	(c)	(d)	(e)
2.391	0.300	0.148	0.220	0.074	0.088

Table 4.2: *Critical bandwidths and p-values for the chondrite data.* $\mathcal{I} = [18, 37]$.

no. of modes (j)	$\hat{h}_{\text{crit},j}$	p-value
2	1.833	0.053
3	0.685	0.710
4	0.478	0.893

The p-values were calculated by drawing 999 resamples of size $n = 22$ from the distribution with density $\hat{f}_{\text{crit}}$, shown in Figure 4.1 (a).

The simulations that were performed in Chapter 3 showed that methods (d) and (e) produced tests with best level accuracy, when the null hypothesis is true, and the most power, when the alternative is true. The p-values 0.074 and 0.088 for methods (d) and (e), respectively, indicate that we should reject the null hypothesis of unimodality. Notice that (d) and (e) reject the null hypothesis at a much lower level than form (b), which is the version of the bandwidth test originally proposed by Silverman (1981). Form (c) does not perform well but that is to be expected for a sample with as few as 22 observations, since version (c) is based entirely on large-sample ideas. One would expect form (c) of the test to outperform form (b) for samples with more than one hundred observations; see Section 3.7 for a discussion of the relative merits of the different versions of the test.

We explained in Section 2.5 that our calibrated versions of the test, forms (c), (d) and (e), only apply when we are testing the null hypothesis of one mode. When we want to test the null hypothesis of j modes, for $j \geq 2$, we need to revert to using

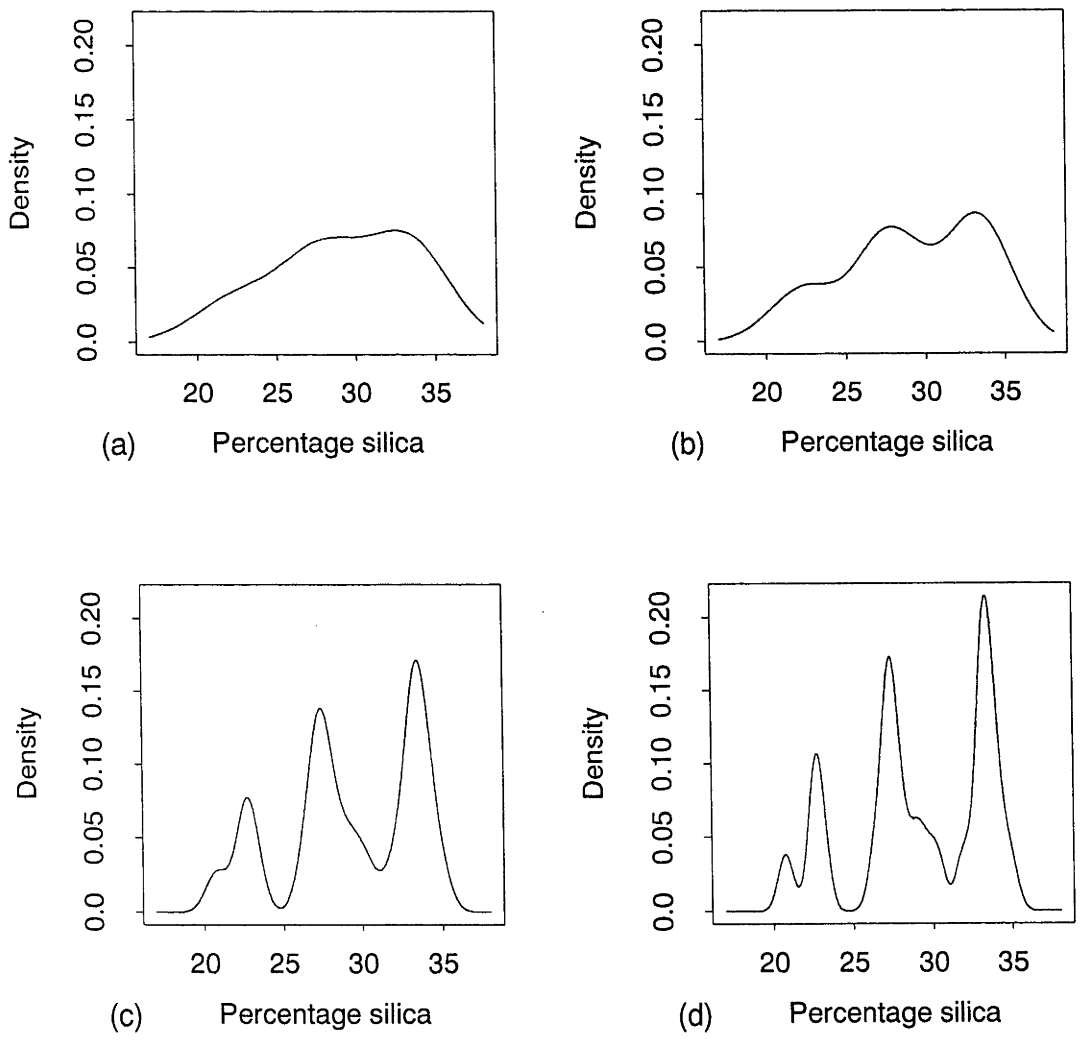


Figure 4.1: *Kernel density estimates based on measurements of the percentages of silica in 22 meteors. The critical bandwidths for (a) one, (b) two, (c) three and (d) four modes have been used.*

Table 4.3: *Critical bandwidth and p-values for testing the null hypothesis of unimodality on the Buffalo snowfall data. Versions (a)-(e) of the test were performed on the interval* $\mathcal{I} = [35, 130]$.

$\hat{h}_{\text{crit}}$	(a)	(b)	(c)	(d)	(e)
7.410	0.676	0.600	0.634	0.552	0.591

form (b) of the test, which is known to be quite conservative.

We continued sequentially, testing for $j = 2, 3$ and 4 modes. The critical bandwidths and the associated p-values are presented in Table 4.2. These values agree closely with those reported by Silverman (1981). This was not quite the case for $j = 1$ where Silverman reports a p-value of 0.08 but we found a value of 0.148. Two explanations for this discrepancy could be the possible simulation error present in Silverman's p-value, since it is only based on 100 bootstrap replications, and the use of different intervals $\mathcal{I}$. The density estimates $\hat{f}_{\text{crit}}$ calculated using the bandwidths $\hat{h}_{\text{crit},j}$, for $j = 1, 2, 3, 4$, are shown in Figure 4.1. There is no doubt here that we should reject the hypothesis of two modes in favour of that of three modes.

4.3 Buffalo snowfall data

The Buffalo snowfall data consist of the measurements of annual snowfall (in inches) in Buffalo, New York, for the 63 winters from 1910-11 to 1972-73. The data set was obtained from Scott (1992). Scott (1980) argues that these data appear to be trimodal while Parzen (1979) has suggested that the evidence leans towards a unimodal density.

The smallest value in the data set, 25.0, is likely to be an outlier so we chose to test for modality over the interval $\mathcal{I} = [35, 130]$ so that this outlying point would not affect the results of the tests. We first tested the null hypothesis of $j = 1$ mode. The critical bandwidth and the p-values for the five forms of the test are reproduced in Table 4.3. With the smallest p-value being 0.552, the test provides no evidence of the presence of more than one mode. The plot of $\hat{f}_{\text{crit}}$ is given in Figure 4.2 (a).

We continued the analysis by testing the null hypothesis of j modes, for $j =$

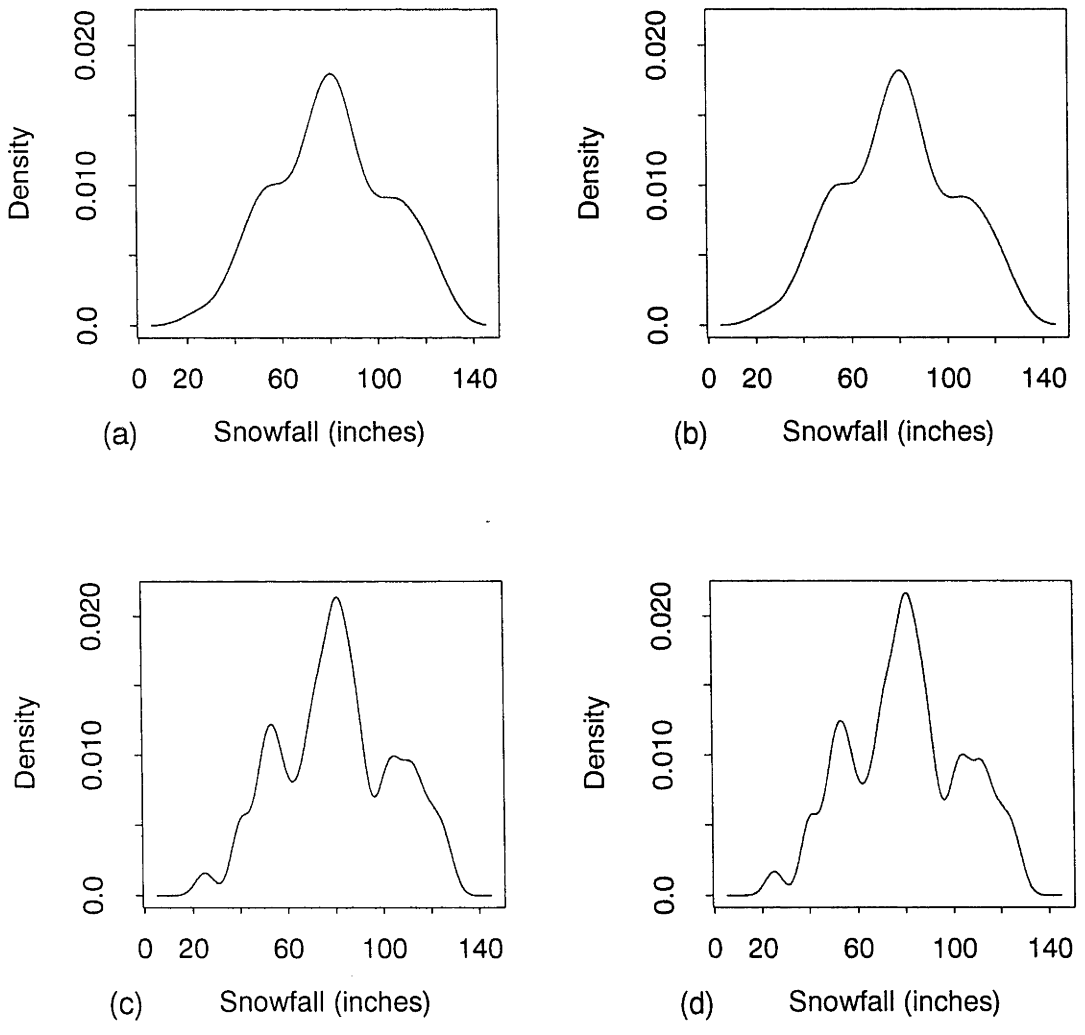


Figure 4.2: *Kernel density estimates for the annual snowfall in Buffalo for 63 winters. The critical bandwidths for (a) one, (b) two, (c) three and (d) four modes have been used.*

Table 4.4: *Critical bandwidths and p-values for the Buffalo snowfall data.* $\mathcal{I} = [35, 130]$.

no. of modes (j)	$\hat{h}_{\text{crit},j}$	p-value
2	7.166	0.161
3	3.940	0.762
4	3.743	0.437

$2, 3, 4$, using form (b) of the test. Forms (c), (d) and (e) can only be used in the case $j = 1$. The p-values and the associated critical bandwidths are given in Table 4.4. The plots of the density estimates with bandwidths $\hat{h}_{\text{crit},j}$, $j = 2, 3, 4$, are given in Figures 4.2 (b)–(d) respectively. If we bear in mind that when we are testing for $j \geq 2$ modes we are using form (b) of the test, which is very conservative, then the p-value of 0.161 for testing $j = 2$ modes might suggest the presence of two or more modes. There is no evidence of more than three modes.

We would argue that the size of the p-value for $j = 2$ is more an artefact of the resampling method than an indication of multimodality. This can be seen by observing the resampling density in Figure 4.2 (b). This density lies on the boundary between bimodal and trimodal densities; it has two modes and a shoulder. However it is very similar to the density pictured in Figure 4.2 (a), which is unimodal. These two densities are so similar that if we were to draw a sample of size $n = 63$ from either of them it would be impossible to determine from which distribution the sample was drawn. Therefore when testing for $j = 2$ modes the bootstrap samples are drawn from a distribution which is very close to being unimodal which results in an artificially low p-value being produced.

Of course, the large p-values that resulted from testing for $j = 1$ mode could be slightly discounted by similarly arguing that in that test the bootstrap resamples were drawn from a distribution that was almost trimodal. It would have to be argued very persuasively indeed, however, for one to consider rejecting the null hypothesis of unimodality, given p-values of around 0.6.

For the sake of thoroughness we tested for the number of modes in the intervals $\mathcal{I} = [35, 95]$ and $\mathcal{I} = [65, 130]$. If we found that the density was unimodal in both

Table 4.5: *Critical bandwidth and p-values for testing the null hypothesis of unimodality on the Buffalo snowfall data. Versions (a)–(e) of the test were performed on the two intervals $\mathcal{I}$ given in the table.*

$\mathcal{I}$	$\hat{h}_{\text{crit}}$	(a)	(b)	(c)	(d)	(e)
$[35, 95]$	7.170	0.429	0.359	0.339	0.293	0.289
$[65, 130]$	7.412	0.471	0.421	0.418	0.348	0.351

these intervals then the density should be unimodal in the interval $\mathcal{I} = [35, 130]$. The critical bandwidths and p-values for testing for $j = 1$ mode are given in Table 4.5 for both intervals. The results strongly support the null hypothesis of unimodality. To complete the analysis we tested for $j = 2$ modes in both the intervals. For the interval $\mathcal{I} = [35, 95]$ the p-value was 0.592 and for the interval $\mathcal{I} = [65, 130]$ the p-value was 0.662.

Consequently we would conclude that the density is unimodal in both these intervals, suggesting that the density is unimodal overall.

4.4 Swiss bank notes data

The Swiss bank notes data set consists of the measurements (in mm) of the height of the bottom margins of 100 forged Swiss bank notes and 100 real Swiss bank notes. These data were obtained from Simonoff (1996) via STATLIB. Simonoff suggests that the distribution of the real notes is unimodal while the distribution of the forged notes is trimodal. The multimodality of the distribution of the forged notes suggests that they were produced in different batches.

We shall consider the data for the real bank notes first. Since the largest observation of 10.4 appeared to be an outlier we chose to test over the interval $\mathcal{I} = [7, 10]$, so that our analysis would not be affected by this point. The usual results for testing for $j = 1$ mode are listed in Table 4.6. The density estimate $\hat{f}_{\text{crit}}$ is shown in Figure 4.3 (a). Methods (d) and (e) are the most reliable forms of the test and they produce p-values of 0.600 and 0.637 respectively, which suggest that the null hypothesis of unimodality is correct. We also tested for $j = 2$ and $j = 3$ modes.

Table 4.6: *Critical bandwidth and p-values for testing the null hypothesis of unimodality on the real Swiss bank notes data. Versions (a)–(e) of the test were performed on the interval* $\mathcal{I} = [7, 10]$.

$\hat{h}_{\text{crit}}$	(a)	(b)	(c)	(d)	(e)
0.1968	0.703	0.654	0.668	0.600	0.637

Table 4.7: *Critical bandwidths and p-values for the real Swiss bank notes data.* $\mathcal{I} = [7, 10]$.

no. of modes (j)	$\hat{h}_{\text{crit},j}$	p-value
2	0.1354	0.706
3	0.1341	0.276

The p-values, given in Table 4.7, are both quite large, so overall we would conclude that the distribution of the height of the bottom margins on the real Swiss bank notes is unimodal.

Turning our attention to the forged bank notes we chose the interval $\mathcal{I} = [7, 13]$ in which to test. The p-values obtained for testing for $j = 1$ mode are shown in Table 4.8. The value of $\hat{h}_{\text{crit}}$ is also given there and the density estimate $\hat{f}_{\text{crit}}$ is plotted in Figure 4.4 (a). This is an example where having a test with good level accuracy is important. The most accurate versions of the test, forms (d) and (e), produce p-values of 0.090 and 0.129 respectively, which are about half the size of the p-value of 0.207 obtained using the usual conservative version of the test, form (b). Here the evidence leans towards rejection of the null hypothesis of unimodality, but it is a borderline case.

We proceeded sequentially, testing the null hypotheses of j modes, for $j = 2, 3, 4$ and 5. The p-values are listed in Table 4.9. They strongly suggest the presence of three modes near 8, 10 and 11.5. The critical bandwidths $\hat{h}_{\text{crit},j}$ are also listed in Table 4.9 and the density estimates using bandwidths $\hat{h}_{\text{crit},j}$, for $j = 1, 2, 3$ and 4 are

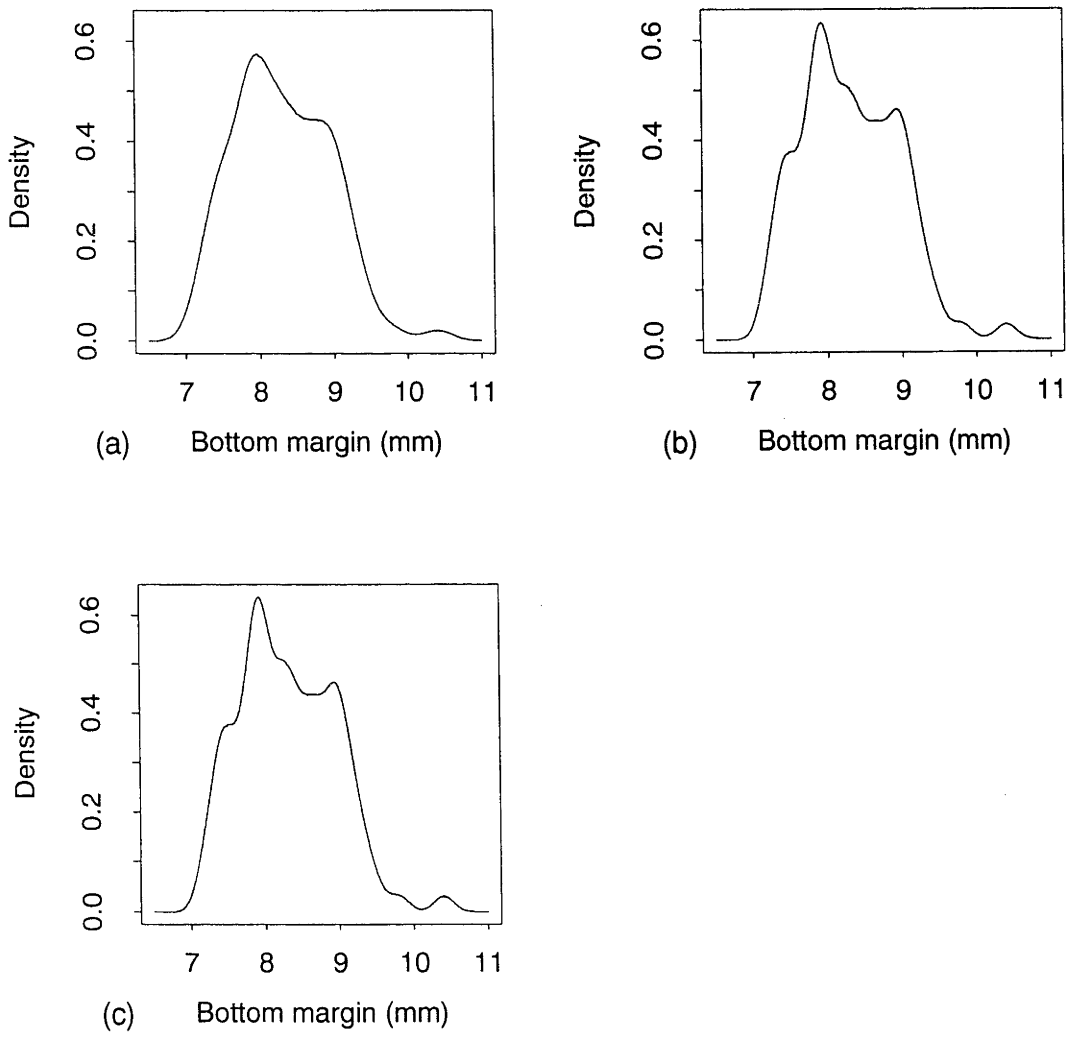


Figure 4.3: *Kernel density estimates for the height of the bottom margins of 100 real Swiss bank notes. The critical bandwidths for (a) one, (b) two and (c) three modes have been used.*

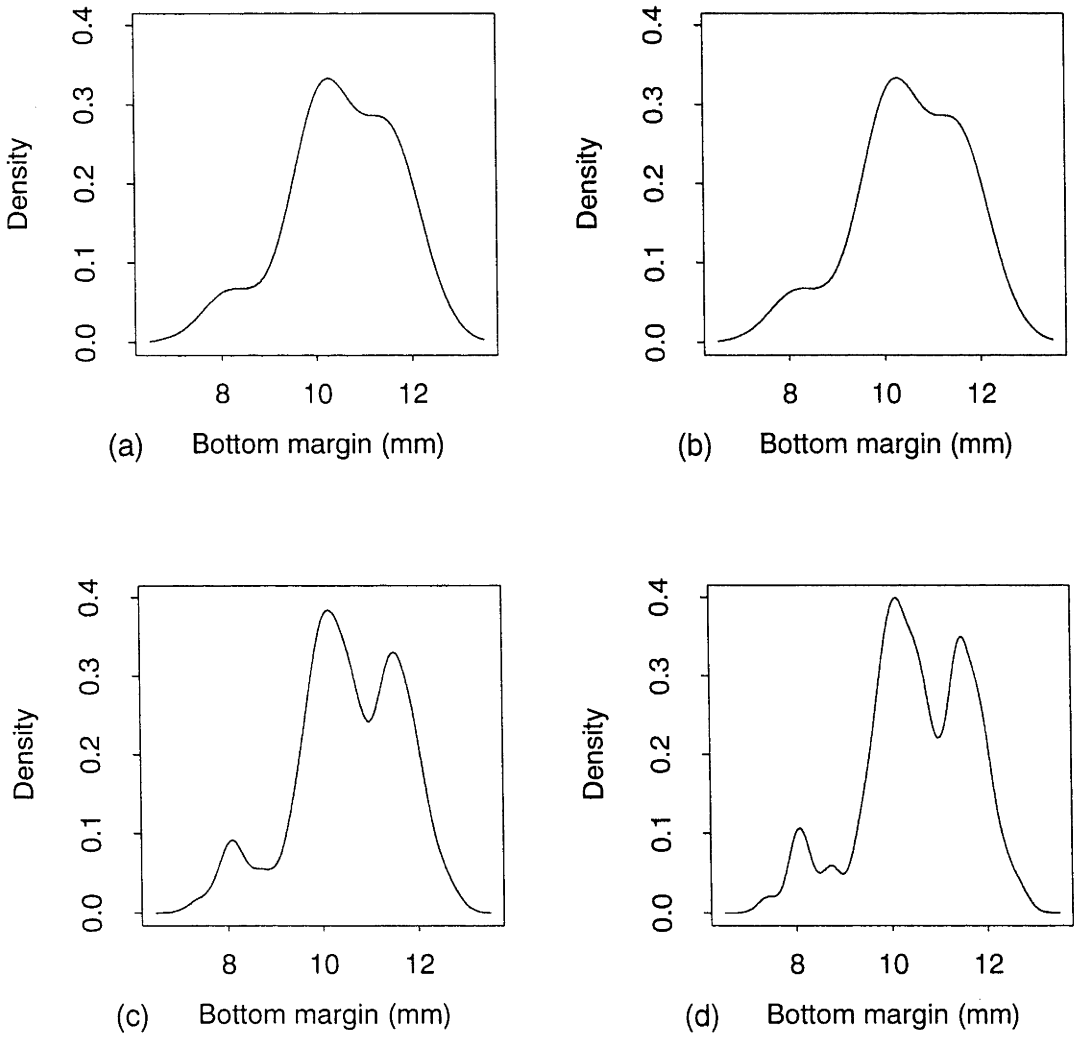


Figure 4.4: *Kernel density estimates for the height of the bottom margins of 100 forged Swiss bank notes. The critical bandwidths for (a) one, (b) two, (c) three and (d) four modes have been used.*

Table 4.8: *Critical bandwidth and p-values for testing the null hypothesis of unimodality on the forged Swiss bank notes data. Versions (a)-(e) of the test were performed on the interval* $\mathcal{I} = [7, 13]$.

$\hat{h}_{\text{crit}}$	(a)	(b)	(c)	(d)	(e)
0.4553	0.326	0.207	0.194	0.090	0.129

Table 4.9: *Critical bandwidths and p-values for the forged Swiss bank notes data.* $\mathcal{I} = [7, 13]$.

no. of modes (j)	$\hat{h}_{\text{crit},j}$	p-value
2	0.4527	0.013
3	0.2535	0.263
4	0.2052	0.244
5	0.1608	0.461

shown in Figure 4.4.

To confirm the existence of the two major modes near 10 and 11.5 we repeated the testing procedure over the shorter interval of $\mathcal{I} = [9.2, 12.5]$. The p-values for testing the null hypothesis of one mode are reproduced in Table 4.10. Forms (d) and (e) of the test produce p-values of 0.046 and 0.053 respectively, which provide marginal evidence that the null hypothesis of unimodality is untrue. Testing for $j = 2$ and $j = 3$ modes yields the p-values presented in Table 4.11. These results confirm the presence of the two modes near 10 and 11.5 and indicate that they are the only two modes in the interval $\mathcal{I} = [9.2, 12.5]$.

In conclusion, we deduce that the distribution of the heights of the bottom margins of the forged Swiss bank notes is trimodal. This would suggest that the forged notes in the sample were produced in different batches.

Table 4.10: *Critical bandwidth and p-values for testing the null hypothesis of unimodality on the forged Swiss bank notes data. Versions (a)–(e) of the test were performed on the interval* $\mathcal{I} = [9.2, 12.5]$.

$\hat{h}_{\text{crit}}$	(a)	(b)	(c)	(d)	(e)
0.4530	0.175	0.124	0.097	0.046	0.053

Table 4.11: *Critical bandwidths and p-values for the forged Swiss bank notes data.* $\mathcal{I} = [9.2, 12.5]$.

no. of modes (j)	$\hat{h}_{\text{crit},j}$	p-value
2	0.1618	0.764
3	0.1489	0.437

4.5 Stamp data

This data set consists of the measurements of the thickness in millimetres of 485 Mexican stamps printed in 1872. It was not unusual for different types of paper to be used in the production of a single stamp issue. It is of interest to collectors to know how many different types of paper were used, and the relative scarcity of each type of paper. Assessing the modality of the distribution of the thickness of the stamps is one method for determining the number of types of paper used. Since these data was first analysed by Izenman and Sommer (1988) they have become a popular benchmarking data set for mode testing techniques. For example, Efron and Tibshirani (1993, Section 16.5) and Minnotte and Scott (1993) have considered the stamp data in the context of mode testing.

We begin with the case $j = 1$ and take $\mathcal{I}$ to be the interval $[0.058, 0.132]$. The critical bandwidth $\hat{h}_{\text{crit}}$ and the p-values, based on 999 bootstrap replications, for the five methods are shown in Table 4.12. Four of the p-values are zero and the fifth is extremely small. Therefore the null hypothesis of unimodality is strongly rejected no matter which form of the test is used. There is overwhelming evidence that the

Table 4.12: *Critical bandwidth and p-values for testing the null hypothesis of unimodality on the stamp data. Versions (a)–(e) of the test were performed on the interval* $\mathcal{I} = [0.058, 0.132]$.

$\hat{h}_{\text{crit}}$	(a)	(b)	(c)	(d)	(e)
0.00672	0.003	0.000	0.000	0.000	0.000

Table 4.13: *Critical bandwidths and p-values for the stamp data.* $\mathcal{I} = [0.058, 0.132]$.

no. of modes (j)	$\hat{h}_{\text{crit},j}$	p-value
2	0.00323	0.298
3	0.00300	0.043
4	0.00283	0.006
5	0.00262	0.000
6	0.00241	0.000
7	0.00148	0.453
8	0.00136	0.235
9	0.00106	0.559

data come from a multimodal distribution.

We continued testing for more than one mode. The critical bandwidths and the corresponding p-values are listed in Table 4.13. The p-values suggest the existence of either two or seven modes. The density estimates $\hat{f}_{\text{crit}}$ constructed using the bandwidths $\hat{h}_{\text{crit},j}$, for $j = 1, 2, \ldots, 9$, are shown in Figure 4.5. Izenman and Sommer (1988) and Efron and Tibshirani (1993) have performed this same analysis. The results reported here agree closely with those of Efron and Tibshirani (except for the p-value for $j = 9$) but are slightly different from those of Izenman and Sommer.

Izenman and Sommer also fitted a mixture of Normals to the data. They decided on a mixture of three components and used maximum likelihood to estimate the

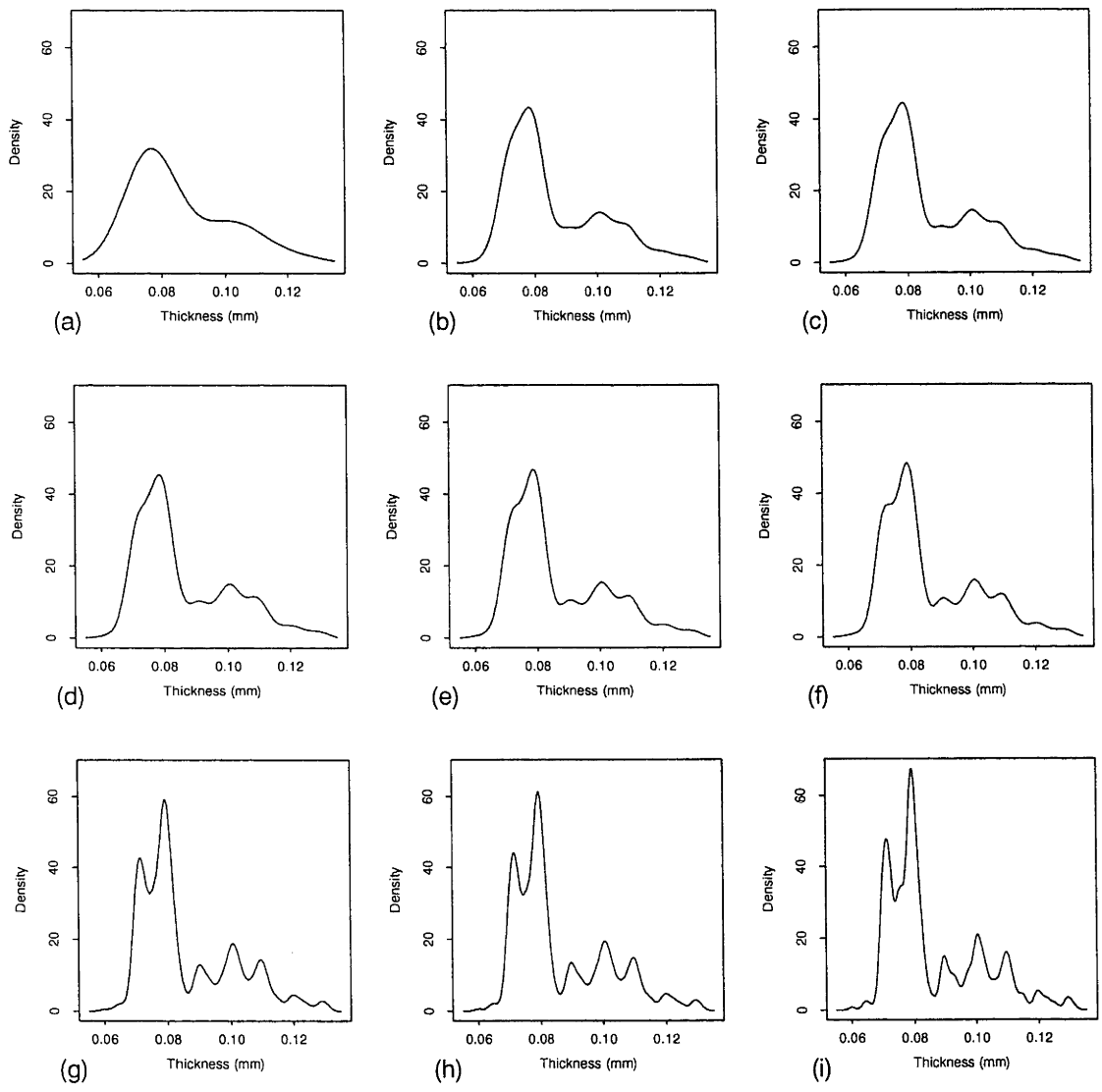


Figure 4.5: *Kernel density estimates for the thickness of 485 Mexican stamps. The critical bandwidths for (a) one, (b) two, (c) three, (d) four, (e) five, (f) six, (g) seven, (h) eight and (i) nine modes have been used.*

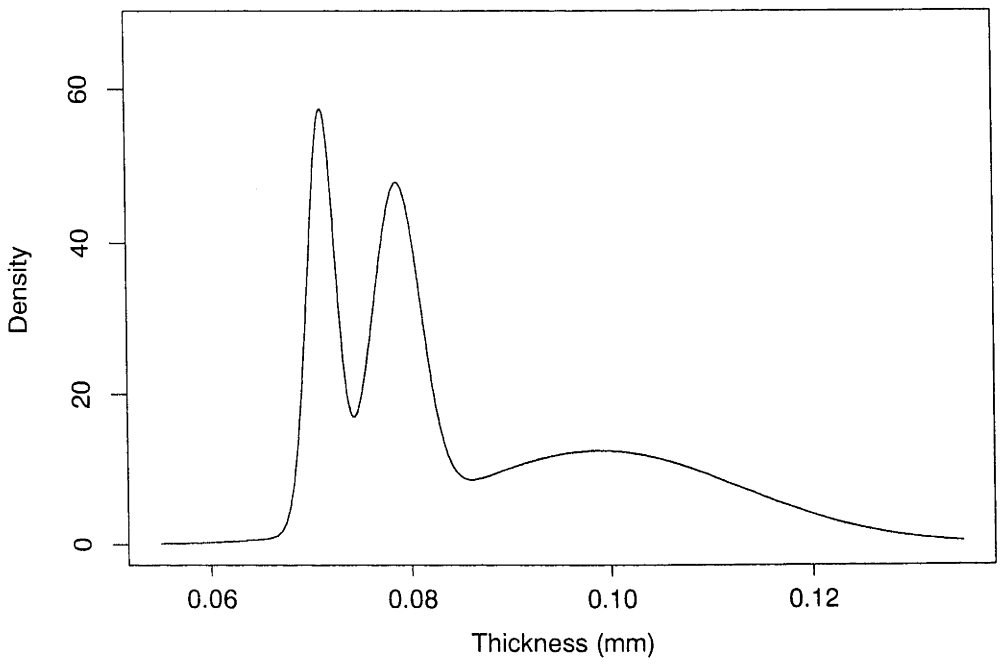


Figure 4.6: *The mixture of three Normals fitted to the stamp data by Izenman and Sommer.*

parameters. The resulting mixture is

$$0.196\,N(0.0712, 0.000002) + 0.367\,N(0.0786, 0.000006) \\ +0.437\,N(0.0989, 0.000197)\,. \qquad (4.1)$$

A graph of this density is given in Figure 4.6. While this model might not be correct it does raise the possibility that a kernel density estimate with a global bandwidth does not estimate the sampling density particularly well. The mixture analysis suggests the presence of three modes, while the bandwidth test indicates either two or seven modes.

The size of the three modes in Figure 4.6 and the separation between the modes would suggest that it is unlikely that the density only has two modes. The high p-value produced by the critical bandwidth test for $j = 2$ would seem to be more of an undesirable consequence of the use of a kernel density estimate with a global bandwidth rather than a serious indication of the presence of exactly two modes.

Table 4.14: *Critical bandwidth and p-values for testing the null hypothesis of unimodality on the stamp data. Versions (a)-(e) of the test were performed on the interval* $\mathcal{I} = [0.088, 0.115]$.

$\hat{h}_{\text{crit}}$	(a)	(b)	(c)	(d)	(e)
0.00323	0.093	0.084	0.027	0.019	0.011

Minnotte (1997) examines the performance of the bandwidth test on a mixture of Normals similar to density (4.1), with two tall narrow modes followed by a short flat mode. He reports that about 90% of the time the test only detects two modes, and that when it does detect three modes it nearly always finds a spurious mode in the region of the flat mode rather than detecting the three correct modes.

The problem here is that the part of the density with the two sharp modes is best estimated with a smaller bandwidth than the part with the flat mode. By the time the bandwidth is small enough to distinguish the two tall modes, a number of spurious modes have appeared in the vicinity of the flat mode. Returning to the stamp data and Figure 4.5, we see that the two prominent modes on the left of the Normal mixture density do not appear as separate modes of the kernel density estimate until the bandwidth is so small that the density estimate has seven modes. Hence, it is possible that f has only three modes and that the bandwidth test indicates seven only because of the use of a global bandwidth.

The most interesting question here is whether f has one mode in the interval $\mathcal{I} = [0.088, 0.115]$, or more, most likely three. This approach allows us to disregard the behaviour of the kernel density estimate outside $\mathcal{I}$, thus incorporating a degree of local adaptivity as in the work of Minnotte and Scott (see Section 1.6) without having to resort to their very conservative techniques for assessing significance. Firstly, we tested for $j = 1$ mode. The value of $\hat{h}_{\text{crit}}$ and the p-values for the five forms of the test are given in Table 4.14. They strongly suggest the presence of multimodality. For method (e) we used a calibration based on a sample size of $n = 200$ since the data set has 153 points located in $\mathcal{I}$. We continued the testing procedure for $j = 2$ and $j = 3$. The critical bandwidths and the corresponding p-values are given in Table 4.15. The null hypothesis of two modes is clearly rejected, and while

Table 4.15: *Critical bandwidths and p-values for the stamp data.* $\mathcal{I} = [0.088, 0.115]$.

no. of modes (j)	$\hat{h}_{\text{crit},j}$	p-value
2	0.00301	0.003
3	0.00106	0.671

Table 4.16: *Critical bandwidth and p-values for testing the null hypothesis of unimodality on the stamp data. Versions (a)–(e) of the test were performed on the interval* $\mathcal{I} = [0.068, 0.084]$.

$\hat{h}_{\text{crit}}$	(a)	(b)	(c)	(d)	(e)
0.00241	0.029	0.023	0.002	0.002	0.001

for $j = 3$ the null is clearly is not. Hence the data strongly indicate that f has three modes in the interval $[0.088, 0.115]$. Overall this would tend to support the hypothesis that f has seven modes.

In their analysis of the stamp data Minnotte and Scott (1993) find evidence of an extra mode near 0.075, that is, between the two tall modes. Again it is possible that the critical bandwidth test misses this mode because of its use of a global bandwidth. To test for the existence of the mode we conducted the critical bandwidth test over the interval $\mathcal{I} = [0.068, 0.084]$. The bandwidth and the usual five p-values are listed in Table 4.16. There is clearly more than one mode in $\mathcal{I}$. The results for $j = 2, 3$ and 4 are given in Table 4.17 but they seem to be inconclusive. The problem here is that the data are rounded to the third decimal place and $\hat{h}_{\text{crit},4}$ is smaller than this rounding error, so the rounding is likely to induce spurious modes. To overcome this difficulty we added a small random perturbation, a Uniform$(-0.0005, 0.0005)$ random variate, to the original data and repeated the testing procedure.

The results for $j = 1$ were virtually unchanged and the outcomes for $j = 2, 3$ and 4 are given in Table 4.18. The p-values suggest either the two modes indicated in the overall fit of seven modes or the three modes found by Minnotte and Scott.

Table 4.17: *Critical bandwidths and p-values for the stamp data.* $\mathcal{I} = [0.068, 0.084]$.

no. of modes (j)	$\hat{h}_{\text{crit},j}$	p-value
2	0.00104	0.116
3	0.00063	0.206
4	0.00056	0.104

Table 4.18: *Critical bandwidths and p-values for the randomly perturbed stamp data.* $\mathcal{I} = [0.058, 0.132]$.

no. of modes (j)	$\hat{h}_{\text{crit},j}$	p-value
2	0.00103	0.136
3	0.00061	0.252
4	0.00048	0.369

Caution would recommend two modes, particularly given the large number of tests we have conducted, but bearing in mind the conservative nature of the test for $j \geq 2$ we cannot entirely rule out three modes on the weight of a p-value of 0.136.

4.6 Galaxy velocity data

The final data set that we analyse consists of measurement of the velocities of 82 galaxies. Roeder (1990) analysed these data using a mixture of Normals and found that the distribution had between three and seven modes. She also applied Silverman's test and found that it indicated the presence of at least three modes.

We shall begin by testing for the number of modes over the interval $\mathcal{I} = [7, 35]$, which includes all the data, and then we shall consider smaller subintervals. The results for testing the null hypothesis of $j = 1$ mode are shown in Table 4.19. A plot of the density estimate $\hat{f}_{\text{crit}}$ is given in Figure 4.7 (a). As usual we have applied the five forms of the test and they all overwhelmingly reject the hypothesis

Table 4.19: *Critical bandwidth and p-values for testing the null hypothesis of unimodality on the galaxy velocity data. Versions (a)–(e) of the test were performed on the interval* $\mathcal{I} = [7, 35]$.

$\hat{h}_{\text{crit}}$	(a)	(b)	(c)	(d)	(e)
3.045	0.006	0.000	0.001	0.000	0.000

Table 4.20: *Critical bandwidths and p-values for the galaxy velocity data.* $\mathcal{I} = [7, 35]$.

no. of modes (j)	$\hat{h}_{\text{crit},j}$	p-value
2	2.481	0.005
3	0.933	0.504
4	0.878	0.234
5	0.726	0.258
6	0.665	0.143
7	0.456	0.720

of unimodality. For form (e) we used a calibration based on a sample size of $n = 100$. Proceeding sequentially we tested the null hypothesis of j modes for $n = 2, 3, \ldots, 7$. The critical bandwidths and the corresponding p-values are listed in Table 4.20. These results closely agree with those of Roeder. Figure 4.7(a)–(g) shows the seven kernel density estimates with the bandwidths $\hat{h}_{\text{crit},j}$, for $j = 1, \ldots, 7$ respectively.

The p-values provide evidence of the existence of at least three modes. The smallish p-value for $j = 6$ and the large one for $j = 7$ suggests that there might be more modes. It is clear that the seven observations less than 12 form one mode and the three observations greater than 30 form another. In the central interval [15,28] there is obviously at least one mode but it is not clear if there are more. As we saw for the stamp data, Silverman's method might just be stopping at three modes because the true density is not well estimated by a kernel density estimate with a global bandwidth. The central interval, where the density rises sharply, would be

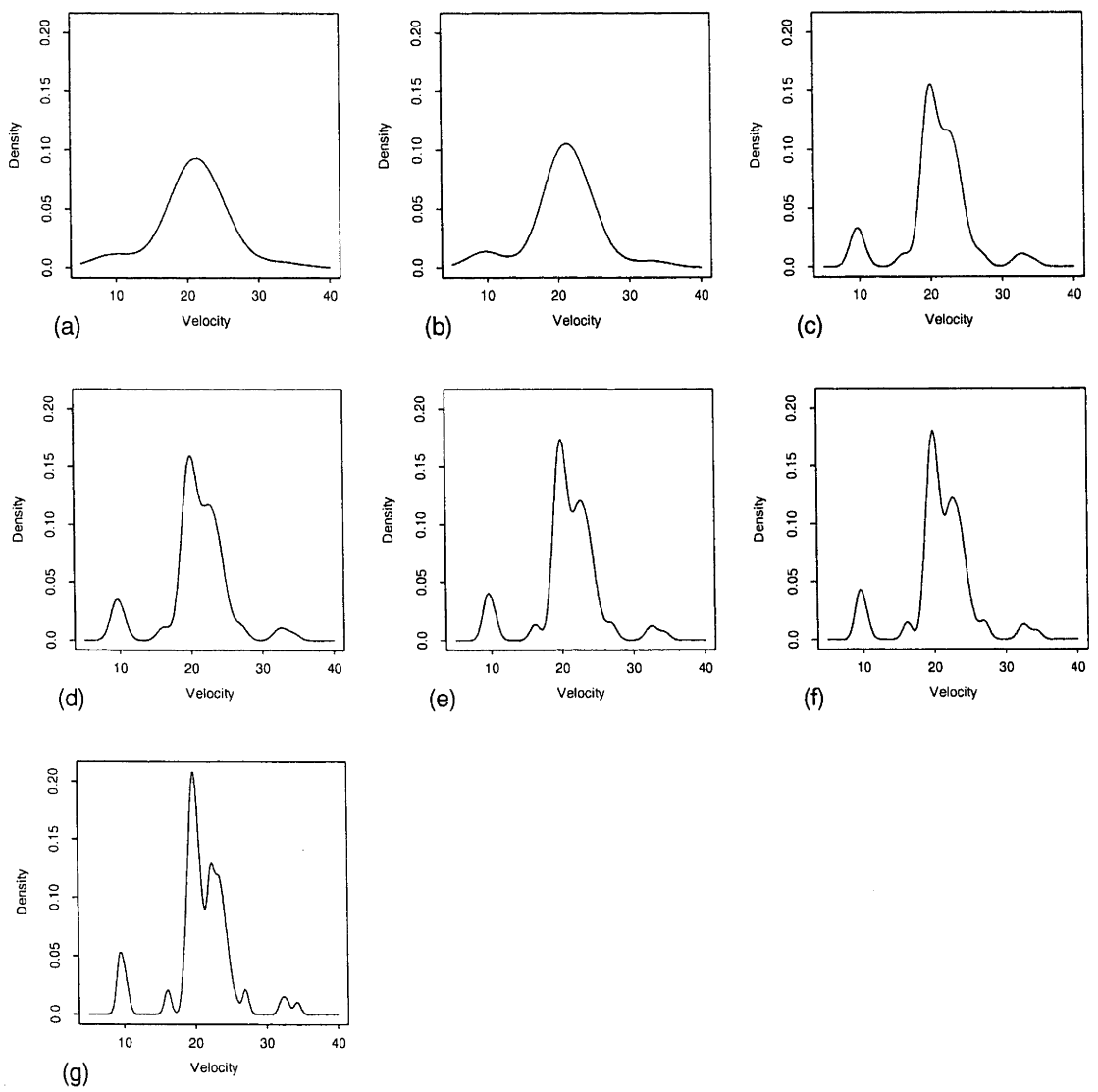


Figure 4.7: *Kernel density estimates for the velocity of 82 galaxies. The critical bandwidths for (a) one, (b) two, (c) three, (d) four, (e) five, (f) six and (g) seven modes have been used.*

Table 4.21: *Critical bandwidth and p-values for testing the null hypothesis of unimodality on the galaxy velocity data. Versions (a)–(e) of the test were performed on the interval* $\mathcal{I} = [15, 28]$.

$\hat{h}_{\text{crit}}$	(a)	(b)	(c)	(d)	(e)
0.936	0.351	0.250	0.330	0.227	0.171

Table 4.22: *Critical bandwidths and p-values for the galaxy velocity data.* $\mathcal{I} = [15, 28]$.

no. of modes (j)	$\hat{h}_{\text{crit},j}$	p-value
2	0.881	0.060
3	0.726	0.074
4	0.449	0.551

better estimated by using a smaller bandwidth than is required for good estimation of the density in the outlying regions, where the density is comparatively flat.

To investigate this we repeated the testing procedure over the central interval $\mathcal{I} = [15, 28]$. The results for $j = 1$ are reproduced in Table 4.21. They support the null hypothesis of unimodality, with even the least conservative version of the test, form (e), only producing a p-value of 0.171. However if we entertain the possibility of multimodality in the interval $\mathcal{I}$ then further testing is in order, which strongly suggests the presence of four modes. This can be seen from the p-values given in Table 4.22.

If there are four modes in the interval $\mathcal{I} = [15, 28]$ then there would be two extremely small modes around 16 and 27 and two large modes around 20 and 22.5. The two extremely small modes are not of much interest since they each contain only two data points; the more interesting question is whether there are two modes at 20 and 22.5 or there is only one mode in this region of high probability density. To answer this question we shortened our interval of interest even further to $\mathcal{I} =$

Table 4.23: *Critical bandwidth and p-values for testing the null hypothesis of unimodality on the galaxy velocity data. Versions (a)–(e) of the test were performed on the interval* $\mathcal{I} = [18, 26]$.

$\hat{h}_{\text{crit}}$	(a)	(b)	(c)	(d)	(e)
0.936	0.269	0.192	0.250	0.158	0.134

$[18, 26]$. The five p-values for testing the null hypothesis of unimodality are given in Table 4.23. Form (e) of the test produces the least conservative p-value, which is 0.134. This would suggest that we accept the null hypothesis that there is only one mode in $\mathcal{I}$, but the result is not entirely conclusive. There are certainly no more than two.

Overall, from the tests that we have conducted, we can conclude that there is one mode located near 10 and another near 33. In the central region there is most likely a single mode at 20 but it is possible that there are two very small modes located near 16 and 27 and a very prominent peak near 20 with a smaller mode on its side near 22.5.

Chapter 5

Mixture data and curve estimation

5.1 Introduction

In the earlier chapters of this thesis we considered the problem of testing for the number of modes in a population. In Chapter 1, particularly in Section 1.8, we discussed the relationship between multimodality and mixture distributions, in particular, the point that the presence of more than one mode is normally interpreted as an indication that a distribution is a mixture made up of a number of components. This chapter is concerned with estimating the means of the components of a mixture. We begin by addressing the problem of estimating the individual means in a two-component mixture in Section 5.2. In the interests of simplicity we restrict our discussion to the estimation of two means but the case for three or more means may be treated similarly. The remainder of the chapter is devoted to the more complex problem of curve estimation from mixture data.

5.2 Estimating two means in a mixture

Let $I_1, \ldots, I_n$ be independent binary random variables which take on the value 1 with probability π, and 0 with probability $1 - \pi$. Let $Y_1, \ldots, Y_n$ be independent random variables which are generated by observing a random variable U_{i1} if $I_i = 1$, and another random variable U_{i2} if $I_i = 0$. Here, $U_{1j}, \ldots, U_{nj}$ are independent and identically distributed as U_j, for $j = 1, 2$. Given $Y_1, \ldots, Y_n$, we wish to estimate the means μ_1 and μ_2 of U_1 and U_2, respectively. In this general form the problem is so ill-posed that it is insoluble. However, if we add the additional condition that the

distributions of U_1 and U_2 are unimodal, then if μ_1 and μ_2 are sufficiently far apart, the bimodality of the distribution of Y will at least be discernible, providing an indication of the general position of the two component means. In this case tests for multimodality may be used to determine the number of components in the mixture.

There are several different settings in which the problem of estimating μ_1 and μ_2 is explicitly solvable. For example, if the distributions of U_1 and U_2 are identical up to changes of location and scale, if they have finite seventh absolute moments, and if their third, fifth and seventh central moments are zero (e.g. if the distributions are symmetric), then, equating the first seven moments of the mixture distribution to the respective empirical moments, we obtain seven equations in seven unknowns (the mixing proportion, the two means, the two variances, and the fourth and sixth moments of the standardised error distribution). Arguing thus we may estimate the two means root-n consistently. More restrictively, if we assume that the error distributions are identical up to changes of location, then μ_1 and μ_2 may be estimated root-n consistently by equating the first five theoretical moments to their empirical counterparts.

Ever since Pearson (1894) used the method of moments to estimate the parameters of a mixture it has been a popular technique for analysing mixtures. This is due to the fact that estimates can be obtained explicitly, and thus this method avoids the computational difficulties that can arise when using the other approaches that we consider in this section. However, owing to the errors associated with estimating high-order moments, moment methods tend not to be asymptotically efficient. Modern computing technology has made the availability of explicit estimates a less important consideration, and today more efficient estimators are generally preferred. The most common alternative techniques are based on minimising measures of distance or maximising likelihood.

Minimum distance methods are usually based on weighted L^2 distance between distribution functions (Choi and Bulgren, 1968), although the distance between densities can be used; see Titterington *et al.* (1985, p. 116) for a list of distance measures that have been suggested by various researchers. Suppose the distribution functions of U_1 and U_2 are respectively

$$G_{\theta,1}(x) = G_\theta\left(\frac{x-\mu_1}{\sigma_1}\right) \quad \text{and} \quad G_{\theta,2}(x) = G_\theta\left(\frac{x-\mu_2}{\sigma_2}\right),$$

where G_θ is the distribution function of a random variable with zero mean and unit

variance, and θ is a vector of unknown parameters. Let $\widehat{F}$ be an estimate of the distribution function $F = \pi\, G_{\theta,1} + (1-\pi)\, G_{\theta,2}$ of Y. For example, $\widehat{F}$ could be the empirical distribution function of the sample of n observations of Y. Then we may generally estimate μ_1 and μ_2 by minimising

$$S(\mu_1, \mu_2, \sigma_1, \sigma_2, \pi, \theta) = \int \left(\widehat{F} - F\right)^2 w \tag{5.1}$$

with respect to all variables on the left-hand side. The values of the estimators $\hat{\mu}_1$ and $\hat{\mu}_2$ are still not completely determined, but ambiguity may generally be removed by insisting that (for example) $\hat{\mu}_1 \le \hat{\mu}_2$.

If G is known only nonparametrically, for example up to smoothness conditions, then it may be approximated via a parametric form based on $m = m(n)$ distinct parameters, which can be included in the argument of S. Arguing thus, and letting m increase sufficiently slowly, we may derive consistent estimators of μ_1 and μ_2. For example, a Gram-Charlier approximation G_θ, not itself a distribution function, would have the form

$$G_\theta = \Phi - \sum_{i=3}^{m} a_i\, H_{i-1}\, \phi\,, \tag{5.2}$$

where ϕ and Φ are the standard normal density and distribution functions, respectively; H_i is the Hermite polynomial, of degree i and satisfying $\int H_i H_j \phi = 0$; $a_i = E\{H_i(W)\}$, with W having distribution G; $m \ge 3$; and $\theta = (a_3, \ldots, a_m)$. See Johnson and Kotz (1970, Section 4.2) for an account of Gram-Charlier expansion. Excluding the terms in $i = 1$ and $i = 2$ from the series at (5.2) serves to ensure that, like G, G_θ is standardised. That is, $\int x^i\, dG_\theta(x) = 0$ or 1 if $i = 1$ or 2 respectively. The $i = 3$ term provides a correction for skewness, the $i = 4$ term adjusts for kurtosis, and so on.

Alternatively, if G_θ is known up to the finite vector θ of parameters, we may instead estimate parameters by maximising

$$L(\mu_1, \mu_2, \sigma_1, \sigma_2, \pi, \theta) = \sum_{i=1}^{n} \log \left\{ \frac{\pi}{\sigma_1}\, \gamma_\theta \left(\frac{Y_i - \mu_1}{\sigma_1} \right) + \frac{1-\pi}{\sigma_2}\, \gamma_\theta \left(\frac{Y_i - \mu_2}{\sigma_2} \right) \right\}, \tag{5.3}$$

where $\gamma_\theta = G'_\theta$ is the density of the standardised error distribution.

In Section 1.8 we observed that there can be difficulties with maximum likelihood methods because the surface given by equation (5.3) is littered with singularities.

In spite of these problems we choose to use maximum likelihood as our method of estimation because in the next section, where we consider the problem of curve estimation from mixture data, we shall see that maximum likelihood methods for mixtures combine neatly with the nonparametric regression concept of local likelihood (Tibshirani and Hastie, 1987).

For the rest of this chapter we shall focus on the important case where $G = G_\theta$ is the standard Normal distribution Φ. In this case the log-likelihood becomes

$$L(\Psi) = \sum_{i=1}^{n} \log \left\{ \frac{\pi}{\sigma_1} \phi\left(\frac{Y_i - \mu_1}{\sigma_1}\right) + \frac{1-\pi}{\sigma_2} \phi\left(\frac{Y_i - \mu_2}{\sigma_2}\right) \right\}, \tag{5.4}$$

where $\Psi = (\mu_1, \mu_2, \sigma_1, \sigma_2, \pi)^T$.

Since the maximiser of equation (5.4) has no explicit solution it is necessary to use a numerical method to compute the maximum likelihood estimates. The EM algorithm (Dempster, Laird and Rubin, 1977) is an appealingly simple and widely-used technique for obtaining maximum likelihood estimates. The EM algorithm relies on the notion of "complete" data. In our situation the complete data are the data we observe, $Y_1, \ldots, Y_n$, plus the unobserved indicator variables $I_1, \ldots, I_n$. If these indicator variables had been observed then the "complete" log-likelihood would be

$$L_C(\Psi) = \sum_{i=1}^{n} \log \left[\left\{ \frac{\pi}{\sigma_1} \phi\left(\frac{Y_i - \mu_1}{\sigma_1}\right) \right\}^{I_i} \left\{ \frac{1-\pi}{\sigma_2} \phi\left(\frac{Y_i - \mu_2}{\sigma_2}\right) \right\}^{1-I_i} \right]. \tag{5.5}$$

The EM algorithm is so-named because it maximises the likelihood by iterating between an expectation step and a maximisation step. The expectation step involves replacing the complete log-likelihood L_C by its expectation, conditional on the observed data and an estimate $\hat{\Psi}$, that is

$$\begin{aligned} Q(\Psi \mid \hat{\Psi}) &= E\left(L_C(\Psi) \mid Y_1, \ldots, Y_n, \hat{\Psi}\right) \\ &= \sum_{i=1}^{n} \left[z_i \log \hat{\pi} + (1 - z_i) \log(1 - \hat{\pi}) + z_i \log \left\{ \frac{1}{\hat{\sigma}_1} \phi\left(\frac{Y_i - \hat{\mu}_1}{\hat{\sigma}_1}\right) \right\} \right. \\ &\qquad \left. + (1 - z_i) \log \left\{ \frac{1}{\hat{\sigma}_2} \phi\left(\frac{Y_i - \hat{\mu}_2}{\hat{\sigma}_2}\right) \right\} \right], \end{aligned}$$

where

$$\begin{aligned} z_i &= E(I_i \,|\, Y_1, \ldots, Y_n, \hat{\Psi}) \\ &= P(I_i = 1 \,|\, Y_i, \hat{\Psi}) \\ &= \left\{ \frac{\hat{\pi}}{\hat{\sigma}_1} \phi\left(\frac{Y_i - \hat{\mu}_1}{\hat{\sigma}_1} \right) \right\} \left\{ \frac{\hat{\pi}}{\hat{\sigma}_1} \phi\left(\frac{Y_i - \hat{\mu}_1}{\hat{\sigma}_1} \right) + \frac{1 - \hat{\pi}}{\hat{\sigma}_2} \phi\left(\frac{Y_i - \hat{\mu}_2}{\hat{\sigma}_2} \right) \right\}^{-1}. \end{aligned}$$

The maximisation step involves finding $\tilde{\Psi}$ that maximises $Q(\Psi|\hat{\Psi})$ as a function of Ψ. Iterating between these two steps results in the following algorithm.

(i) Start with some initial estimate $\Psi^0 = \left\{ \hat{\mu}_1^{(0)}, \hat{\mu}_2^{(0)}, \hat{\sigma}_1^{(0)}, \hat{\sigma}_2^{(0)}, \hat{\pi}^{(0)} \right\}$.

(ii) Let $p = 0$.

(iii) For $i = 1, \ldots, n$

$$\begin{aligned} z_i = \left[\frac{\hat{\pi}^{(p)}}{\hat{\sigma}_1^{(p)}} \phi\left\{ \frac{Y_i - \hat{\mu}_1^{(p)}}{\hat{\sigma}_1^{(p)}} \right\} \right] \Bigg[\frac{\hat{\pi}^{(p)}}{\hat{\sigma}_1^{(p)}} \phi\left\{ \frac{Y_i - \hat{\mu}_1^{(p)}}{\hat{\sigma}_1^{(p)}} \right\} \\ + \frac{1 - \hat{\pi}^{(p)}}{\hat{\sigma}_2^{(p)}} \phi\left\{ \frac{Y_i - \hat{\mu}_2^{(p)}}{\hat{\sigma}_2^{(p)}} \right\} \Bigg]^{-1}. \end{aligned}$$

(iv) Update $p = p + 1$.

(v) Update the parameter estimates:

$$\begin{aligned} \hat{\pi}^{(p)} &= n^{-1} \sum_{i=1}^{n} z_i, \\ \hat{\mu}_1^{(p)} &= \left(n\hat{\pi}^{(p)} \right)^{-1} \sum_{i=1}^{n} z_i Y_i, \\ \hat{\mu}_2^{(p)} &= \left(n - n\hat{\pi}^{(p)} \right)^{-1} \sum_{i=1}^{n} (1 - z_i) Y_i, \\ \hat{\sigma}_1^{2(p)} &= \left(n\hat{\pi}^{(p)} \right)^{-1} \sum_{i=1}^{n} z_i \left(Y_i - \hat{\mu}_1^{(p)} \right)^2, \\ \hat{\sigma}_2^{2(p)} &= \left(n - n\hat{\pi}^{(p)} \right)^{-1} \sum_{i=1}^{n} (1 - z_i) \left(Y_i - \hat{\mu}_2^{(p)} \right)^2. \end{aligned}$$

(vi) Repeat from step (iii) until convergence is obtained.

Dempster, Laird and Rubin (1977) and Wu (1983) studied the conditions under which this algorithm converges to a local maximum of the log-likelihood (5.4).

5.3 Nonparametric curve estimation with mixture data

In this section we adapt the techniques, which were described in the previous section for estimating the means in a mixture, to the situation of estimating the components in a mixture of two smooth regression curves. First we describe a model for generating independent and identically distributed data pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$. Suppose $X_1, \ldots, X_n$ are independent random variables with a common continuous distribution. Given the X_i's, let $I_1, \ldots, I_n$ be independent binary random variables taking the value 1 with probability π, and 0 with probability $1-\pi$. Let $\epsilon_1, \ldots, \epsilon_n$ be independent random variables, independent also of the X_i's and the I_i's, and having a standard Normal distribution. Conditional on X_i and I_i, let

$$Y_i = \begin{cases} m_1(X_i) + \sigma_1 \epsilon_i & \text{if } I_i = 1 \\ m_2(X_i) + \sigma_2 \epsilon_i & \text{if } I_i = 0, \end{cases}$$

where m_1 and m_2 are twice-differentiable functions, and σ_1 and σ_2 are positive constants.

The assumption of homoscedastic errors in each component may be relaxed by modelling the way in which the variances change with location, although doing so uses degrees of freedom which, in this relatively complex problem, are sometimes better employed to estimate other quantities. Similarly, we could allow the probability π to vary with location. However, allowing π, σ_1 and σ_2 to vary with location will increase the chance that numerical estimation problems, associated with the presence of singularities in mixture likelihood surfaces, will be encountered. See Section 1.8 for discussion of these singularities. The assumption that the errors ϵ_i are Normally distributed may also be relaxed, and departures from Normality may be modelled by a Gram-Charlier approximation, as they were in Section 5.2.

In this section we wish to develop methods for estimating the smooth curves m_1 and m_2. We choose to use local linear methods for estimating these curves since, in standard regression settings, it is known that local polynomial kernel estimators have many desirable features. These include satisfactory boundary behaviour,

adaptability to various types of design (Fan, 1992), near optimum minimax efficiency among the class of linear smoothers (Fan, 1993), and the existence of fast, efficient algorithms which can compute local polynomial fits in $O(n)$ operations (Fan and Marron, 1994). For overviews of local polynomial modelling see Wand and Jones (1995, Chapter 5) and Fan and Gijbels (1996).

To estimate the curves m_1 and m_2 we adapt the mixture log-likelihood given by equation (5.4) to the local linear framework, to obtain

$$L_x(\alpha_1,\alpha_2,\beta_1,\beta_2) = \sum_{i=1}^{n} K\left(\frac{X_i-x}{h}\right)\log\left[\frac{\pi}{\sigma_1}\phi\left\{\frac{Y_i-\alpha_1-\beta_1(X_i-x)}{\sigma_1}\right\} + \frac{1-\pi}{\sigma_2}\phi\left\{\frac{Y_i-\alpha_2-\beta_2(X_i-x)}{\sigma_2}\right\}\right], \quad (5.6)$$

where K is a kernel function, which we take to be the standard Normal density, and h is a bandwidth which controls the amount of smoothing. If $\hat{\alpha}_{x1}, \hat{\alpha}_{x2}, \hat{\beta}_{x1}$ and $\hat{\beta}_{x2}$ are the estimators of $\alpha_1, \alpha_2, \beta_1$ and β_2, derived by minimising L_x, then our estimators of $m_1(x)$ and $m_2(x)$ are $\hat{\alpha}_{x1}$ and $\hat{\alpha}_{x2}$ respectively.

Once we have obtained estimates $\widehat{m}_1(\cdot\,;h)$ and $\widehat{m}_2(\cdot\,;h)$ for $m_1(\cdot)$ and $m_2(\cdot)$ we may obtain estimates of σ_1, σ_2 and π by maximising the log-likelihood

$$L(\sigma_1,\sigma_2,\pi) = \sum_{i=1}^{n}\log\left[\frac{\pi}{\sigma_1}\phi\left\{\frac{Y_i-\widehat{m}_1(X_i;h)}{\sigma_1}\right\} + \frac{1-\pi}{\sigma_2}\phi\left\{\frac{Y_i-\widehat{m}_2(X_i;h)}{\sigma_2}\right\}\right]. \quad (5.7)$$

To maximise either the local log-likelihood at (5.6) or the log-likelihood at (5.7), we may use the EM algorithm as we did in Section 5.2. To maximise (5.6) and (5.7) simultaneously we propose the following procedure, which is an EM-style algorithm that alternates between taking an EM-iteration towards maximising equation (5.6) and taking an EM-iteration towards maximising equation (5.7).

(i) Start with some initial estimates $\widehat{m}_1^{(0)}(\cdot\,;h_1), \widehat{m}_2^{(0)}(\cdot\,;h_2), \hat{\sigma}_1^{(0)}, \hat{\sigma}_2^{(0)}, \hat{\pi}^{(0)}$.

(ii) Let $p = 0$.

(iii) For $i = 1, \ldots, n$,

$$z_i = \left[\frac{\hat{\pi}^{(p)}}{\hat{\sigma}_1^{(p)}}\phi\left\{\frac{Y_i-\widehat{m}_1^{(p)}(X_i;h_1)}{\hat{\sigma}_1^{(p)}}\right\}\right]\left[\frac{\hat{\pi}^{(p)}}{\hat{\sigma}_1^{(p)}}\phi\left\{\frac{Y_i-\widehat{m}_1^{(p)}(X_i;h_1)}{\hat{\sigma}_1^{(p)}}\right\} + \frac{1-\hat{\pi}^{(p)}}{\hat{\sigma}_2^{(p)}}\phi\left\{\frac{Y_i-\widehat{m}_2^{(p)}(X_i;h_2)}{\hat{\sigma}_2^{(p)}}\right\}\right]^{-1}. \quad (5.8)$$

(iv) Update $p = p + 1$.

(v) For a grid of values of x compute:

$$\widehat{m}_1^{(p)}(x; h_1) = \frac{\hat{s}_{21}(x; h_1)\hat{t}_{01}(x; h_1) - \hat{s}_{11}(x; h_1)\hat{t}_{11}(x; h_1)}{\hat{s}_{21}(x; h_1)\hat{s}_{01}(x; h_1) - \hat{s}_{11}(x; h_1)^2},$$

where

$$\hat{s}_{r1}(x; h_1) = \sum_{i=1}^{n} z_i (X_i - x)^r K\left(\frac{X_i - x}{h_1}\right) \quad \text{and} \tag{5.9}$$

$$\hat{t}_{r1}(x; h_1) = \sum_{i=1}^{n} z_i (X_i - x)^r K\left(\frac{X_i - x}{h_1}\right) Y_i\,. \tag{5.10}$$

Similarly,

$$\widehat{m}_2^{(p)}(x; h_2) = \frac{\hat{s}_{22}(x; h_2)\hat{t}_{02}(x; h_2) - \hat{s}_{12}(x; h_2)\hat{t}_{12}(x; h_2)}{\hat{s}_{22}(x; h_2)\hat{s}_{02}(x; h_2) - \hat{s}_{12}(x; h_2)^2},$$

where

$$\hat{s}_{r2}(x; h_2) = \sum_{i=1}^{n} (1 - z_i)(X_i - x)^r K\left(\frac{X_i - x}{h_2}\right) \quad \text{and} \tag{5.11}$$

$$\hat{t}_{r2}(x; h_2) = \sum_{i=1}^{n} (1 - z_i)(X_i - x)^r K\left(\frac{X_i - x}{h_2}\right) Y_i\,. \tag{5.12}$$

(vi) Update the global parameters

$$\hat{\pi}^{(p)} = n^{-1} \sum_{i=1}^{n} z_i\,,$$

$$\hat{\sigma}_1^{2(p)} = \left\{n\hat{\pi}^{(p)}\right\}^{-1} \sum_{i=1}^{n} z_i \left\{Y_i - \widehat{m}_1^{(p)}(X_i; h_1)\right\}^2 \quad \text{and}$$

$$\hat{\sigma}_2^{2(p)} = \left\{n - n\hat{\pi}^{(p)}\right\}^{-1} \sum_{i=1}^{n} (1 - z_i) \left\{Y_i - \widehat{m}_2^{(p)}(X_i; h_2)\right\}^2.$$

(vii) Repeat from step (iii) until convergence is obtained.

We do not study the convergence properties of this algorithm or the theoretical properties of the estimates which it produces; rather we present it as a procedure

which has both natural statistical appeal and practical utility. The practical effectiveness of the algorithm is demonstrated in the simulation study in Section 5.6. The combination of EM iterations and smoothing methods was considered by Dempster, Laird and Rubin (1977), and they demonstrated that the EM algorithm will converge to at least a local maximum of a penalised likelihood. Given the close relationship between smoothing splines, which are maximum penalised likelihood estimates, and kernel methods (Silverman, 1984), it seems reasonable to conjecture that, since the EM algorithm converges to an appropriate solution for a smoothing method based on a penalty function, it will also converge to an appropriate estimate when the smoothing is conducted via a kernel method. Silverman, Jones, Wilson and Nychka (1990) studied a problem which also combined an EM iteration with a smoothing step and they found that their algorithm converged relatively quickly to good estimates.

The choice of the bandwidths h_1 and h_2 is a very important aspect of the algorithm. In Section 5.4 we examine this problem and propose the use of the bandwidths $\hat{h}_{1,\mathrm{DPI}}$ and $\hat{h}_{2,\mathrm{DPI}}$, which are defined at equations (5.17) and (5.18) respectively.

The naive estimators $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ for the component variances may be improved on, at least in theoretical terms, by the estimators $\hat{\sigma}_1^2(\hat{\lambda}_{1,\mathrm{MSE}})$ and $\hat{\sigma}_2^2(\hat{\lambda}_{2,\mathrm{MSE}})$, which are defined at equations (5.13) and (5.14). The bandwidths $\hat{\lambda}_{1,\mathrm{MSE}}$ and $\hat{\lambda}_{2,\mathrm{MSE}}$ are given by equations (5.15) and (5.16). The advantages of these estimators are mentioned in Section 5.4, but in practical terms little will be lost by using the naive estimators.

Efficient computational implementation of kernel methods relies on binning methods. These were first proposed for the kernel density estimator by Silverman (1982). Fan and Marron (1994) described fast implementations of local linear regression estimators. They demonstrated that their methods were faster than direct implementations by factors up into the hundreds. Jones (1989) and Hall and Wand (1996) studied the accuracy of binned estimators.

Matt Wand has written S-PLUS functions for efficiently computing local linear estimators and for implementing the bandwidth selector proposed by Ruppert, Sheather and Wand (1995). These routines accompany the monograph on kernel smoothing by Wand and Jones (1995), and are freely available from the World Wide Web, at the URL address

`http://www.biostat.harvard.edu/~mwand/software.html`.

We use these routines throughout the remainder of this chapter for calculating local

linear estimators and implementing the bandwidth selector of Ruppert *et al.* (1995). We have developed modified versions of these routines to implement our algorithm for estimating smooth curves from mixture data, which was given above, and to implement the mixture bandwidth selector that we propose in Section 5.4.

Like most iterative algorithms, the algorithm for estimating the curves from mixture data requires the specification of initial parameter estimates. In practice this normally means the specification of starting values for the z_i's, and from these values the initial parameter estimates may be obtained by following steps (v) and (vi) of the algorithm. One suggestion for obtaining these initial values for the z_i's is to fit a local linear estimator to the data and set z_i to 0 for those points (X_i, Y_i) that lie above the estimated curve, and set z_i to 1 for the remaining points. This is an excellent approach if the curves m_1 and m_2 do not intersect; it does not work so well for curves that do. Exploratory data analysis techniques can be useful. In particular, we found calculating bivariate kernel density estimators for the data, and noting the locations of the modes of the conditional densities, to be quite helpful. Random partitions of the data, where the z_i's are randomly assigned the value 0 or 1, are a good automatic starting point but it can take a number of starts to find the best estimates. When additional information about the curves is available more subjective starting values may be chosen.

The use of different starting values can lead to different estimates for the curves. It would be useful to have some criterion for ranking the different estimates. A natural candidate for this criterion would be one based on an estimate of the mean squared errors of the estimates $\widehat{m}_1$ and $\widehat{m}_2$, such as the estimate used by Fan and Gijbels (1995) and Fan, Gijbels, Hu and Huang (1996). We do not further develop this idea in this thesis.

One problem that we have noted with our estimation procedure is that, if the two curves m_1 and m_2 are quite similar, the algorithm can produce curves $\widehat{m}_1$ and $\widehat{m}_2$ which are poor estimates of m_1 and m_2 respectively. The algorithm might produce a curve $\widehat{m}_1$ which estimates both m_1 and m_2 well in one region and a curve $\widehat{m}_2$ which estimates them both well in a different region. This is the sort of result we would expect if we tried to fit a mixture of two curves to data generated from a single smooth curve. To circumvent this problem we recommend fitting a local linear estimator $\widehat{m}(\cdot\,; h)$ to the data first, then calculating the residuals $r_i = Y_i - \widehat{m}(X_i)$, and then applying the algorithm to the pairs (X_i, r_i), rather than the original data

(X_i, Y_i), to estimate $\widehat{m}_1$ and $\widehat{m}_2$. Then the initial estimate $\widehat{m}(x; h)$ is added to the curves $\widehat{m}_1$ and $\widehat{m}_2$ to obtain the estimates for the original data. The common estimate $\widehat{m}(\cdot\,; h)$ models the features that are common to both m_1 and m_2; by subtracting this curve from the data we remove the features that are common to both m_1 and m_2, and emphasise the differences between them. The residuals will be smoother and better separated than the original data and hence it will be easier to estimate $\widehat{m}_1$ and $\widehat{m}_2$. For all the examples in this chapter we apply this approach before using the curve fitting algorithm.

5.4 Bandwidth selection

As with any kernel smoothing method, the practical performance of the algorithm given in the previous section is heavily dependent on choice of the bandwidth h. If h is chosen too small, then insufficient smoothing is done and the estimated curves are too rough and contain spurious features that are artefacts of the sampling process. If excessive smoothing is done, then important features of the curves are smoothed away. This dilemma can be considered in terms of the trade-off between bias and variance; for small bandwidths the estimated curves have low bias and high variance and for large bandwidths they have high bias and low variance.

Visual choice of the bandwidth can be a very useful way to analyse data but it requires a degree of expertise or prior knowledge, and it introduces an element of subjectivity. In many situations objective, automatic bandwidth selectors are desirable and convenient, if not essential. Up until a decade ago smoothing parameter methods tended to be based on cross-validation, but Härdle, Hall and Marron (1988) showed that such techniques exhibit poor asymptotic and practical performance.

More recent bandwidth selection rules have employed "plug-in" ideas. These methods involve estimating functionals that appear in formulas for asymptotically optimal bandwidths and then plugging the estimates into those formulas. Ruppert, Sheather and Wand (1995) proposed a simple plug-in bandwidth selector for local linear regression and showed that it possessed attractive theoretical and practical properties. We shall describe their method and then adapt it to the problem of estimating curves from mixture data.

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a set of independent and identically distributed data pairs and let the X_i's share a common density f with support confined to a

compact interval $\mathcal{S} = [a, b]$. Let

$$Y_i = m(X_i) + \sigma\epsilon_i, \qquad i = 1, \ldots, n,$$

where $\epsilon_1, \ldots, \epsilon_n$ are independent and identically distributed random variables, independent also of the X_i's, with zero mean, unit variance and finite fourth moment. We shall use $\widehat{m}(x; h)$ to denote the local linear estimator, with bandwidth h, of $m(x)$.

Ruppert *et al.* (1995) chose the conditional weighted mean integrated squared error (MISE) of $\widehat{m}(\cdot\,; h)$, given by

$$\text{MISE}\left\{\widehat{m}(\cdot\,; h) \,|\, X_1, \ldots, X_n\right\} = E\left[\int \left\{\widehat{m}(x; h) - m(x)\right\}^2 f(x)\, dx \,\middle|\, X_1, \ldots, X_n\right],$$

as a convenient criterion against which to select an asymptotically optimal bandwidth. The asymptotic approximation of MISE is

$$\begin{aligned}
&\text{MISE}\left\{\widehat{m}(\cdot\,; h) \,|\, X_1, \ldots, X_n\right\} \\
&\quad = n^{-1}h^{-1}R(K)(b-a)\sigma^2 + \frac{1}{4}h^4\mu_2(K)^2 \int m''(x)^2 f(x)\, dx \\
&\qquad + o_p(n^{-1}h^{-1} + h^4),
\end{aligned}$$

where $R(K) = \int K(u)^2\, du$ and $\mu_2(K) = \int u^2 K(u)\, du$. It is assumed that these integrals converge and $\int$ is taken to mean integration over the whole real line. Thus, the MISE-optimal bandwidth has the asymptotic approximation

$$h_{\text{MISE}} \approx C_1(K)\left\{\frac{(b-a)\sigma^2}{\theta_{22} n}\right\}^{1/5},$$

where $C_1(K) = \{R(K)/\mu_2(K)^2\}^{1/5}$ and $\theta_{rs} = \int m^{(r)}(x) m^{(s)}(x) f(x)\, dx$, for $r, s \geq 0$ and $r + s$ even. Plug-in bandwidth selectors rely on replacing σ^2 and θ_{22} in this approximation for h_{MISE} by estimators.

The estimators Ruppert *et al.* (1995) used were

$$\hat{\theta}_{22}(g) = n^{-1} \sum_{i=1}^{n} \widehat{m}''(X_i; g)\widehat{m}''(X_i; g)$$

and

$$\hat{\sigma}^2(\lambda) = \nu^{-1} \sum_{i=1}^{n} \left\{Y_i - \widehat{m}(X_i; \lambda)\right\}^2.$$

Here, $\nu = n - 2\sum_i w_{ii} + \sum\sum_{ij} w_{ij}^2$, where

$$w_{ij} = \frac{\{\hat{s}_2(X_i;\lambda) - \hat{s}_1(X_i;\lambda)(X_i - x)\}K\{(X_i - x)/\lambda\}}{\hat{s}_2(X_i;\lambda)\hat{s}_0(X_i;\lambda) - \hat{s}_1(X_i;\lambda)^2}$$

and

$$\hat{s}_r(x;\lambda) = \sum_{i=1}^{n} (X_i - x)^r K\left(\frac{X_i - x}{\lambda}\right).$$

The estimator $\hat{\theta}_{22}(g)$ is the natural kernel-type estimator for θ_{22} but the estimator of σ^2 perhaps requires a little explanation. It is motivated by the fact that the residual sum of squares satisfies

$$E\left[\sum_{i=1}^{n} \{Y_i - \widehat{m}(X_i;\lambda)\}^2 \,\middle|\, X_1, \ldots, X_n\right] = \nu\sigma^2,$$

whenever m is a straight line.

The use of $\hat{\theta}_{22}(g)$ and $\sigma^2(\lambda)$ requires selection of the bandwidths g and λ. Ruppert *et al.* (1995) derived asymptotically MSE-optimal choices

$$g_{\text{MSE}} = C_2(K)\left\{\frac{\sigma^2(b-a)}{|\hat{\theta}_{24}|n}\right\}^{1/7}$$

and

$$\lambda_{\text{MSE}} = C_3(K)\left\{\frac{\sigma^4(b-a)}{\hat{\theta}_{22}^2 n^2}\right\}^{1/9},$$

where $C_2(K) = \{3/(8\sqrt{\pi})\}^{1/7}$ if $\theta_{24} < 0$ and $\{15/(16\sqrt{\pi})\}^{1/7}$ if $\theta_{24} > 0$, and $C_3(K) = \{4(1/2 + 2\sqrt{2} - 4\sqrt{3}/3)/\sqrt{2\pi}\}^{1/9}$.

Of course, to compute these bandwidths we need estimates of σ^2, θ_{22} and θ_{24}. These could also be estimated by kernel estimators but this would lead to further bandwidth selection problems, and would require more functionals to be estimated. At some stage a rule-of-thumb estimate of m is needed. Ruppert *et al.* (1995) used a fast and simple estimate, proposed by Härdle and Marron (1995), to obtain an initial estimate of m. This estimate involves partitioning the range of the X data into N blocks and fitting a quartic by ordinary least-squares to each block. The partition is formed by dividing the data into equal sized subsamples. Let $\widehat{m}^Q(x;N)$ be the estimator obtained by this method.

Ruppert *et al.* (1995) used Mallows' C_p (Mallows, 1973) to provide a rule for choosing N. For blocked quartic fits, this involved choosing $\widehat{N}$ from the set $\{1, 2, \ldots, N_{\max}\}$ to minimise

$$C_p(N) = \mathrm{RSS}\,(N)/\{\mathrm{RSS}\,(N_{\max})/(n - 5N_{\max})\} - (n - 10N),$$

where $\mathrm{RSS}\,(N)$ is the residual sum of squares for $\widehat{m}^Q(x; N)$. They suggested taking $N_{\max} = \max\{\min(n/20, N^*), 1\}$ with $N^* = 5$.

Once we have the estimate $\widehat{m}^Q(x; N)$ the "blocked quartic estimator" for σ^2 is

$$\hat{\sigma}_Q^2(N) = (n - 5N)^{-1} \sum_{i=1}^{n} \{Y_i - \widehat{m}^Q(X_i; N)\}^2.$$

Similarly, the blocked quartic estimator for θ_{rs}, with $\max(r, s) \leq 4$, is

$$\hat{\theta}_{rs}^Q(N) = n^{-1} \sum_{i=1}^{n} (\widehat{m}^Q)^{(r)}(X_i; N)(\widehat{m}^Q)^{(s)}(X_i; N)\,.$$

Local linear estimators of higher derivatives can be extremely variable near the boundary so it can be beneficial to truncate the data within $100\alpha\%$ of the boundaries, for some small value of α, when estimating θ_{rs}. In the case where the X_i's are supported on $[a, b]$, this involves the replacement of $\hat{\theta}_{rs}(g)$ by

$$\hat{\theta}_{rs}^{\alpha}(g) = n^{-1} \sum_{i=1}^{n} \widehat{m}^{(r)}(X_i; g)\widehat{m}^{(s)}(X_i; g) 1_{\{(1-\alpha)a+\alpha b < X_i < \alpha a+(1-\alpha)b\}}.$$

In this chapter we take $\alpha = 0.05$.

The direct plug-in bandwidth selector $\hat{h}_{\mathrm{DPI}}$ proposed by Ruppert *et al.* (1995) is computed through the following algorithm.

(i) Find $\hat{\theta}_{24}^Q(\widehat{N})$ and $\hat{\sigma}_Q^2(\widehat{N})$ based on a blocked quartic fit with $\widehat{N}$ chosen by Mallows' C_p.

(ii) Estimate θ_{22} using $\hat{\theta}_{22}^{.05}(\hat{g}_{\mathrm{MSE}})$, where

$$\hat{g}_{\mathrm{MSE}} = C_2(K) \left\{ \frac{\sigma_Q^2(\widehat{N})(b - a)}{|\hat{\theta}_{24}^Q(\widehat{N})|n} \right\}^{1/7}$$

and estimate σ^2 using $\hat{\sigma}^2(\hat{\lambda}_{\mathrm{MSE}})$, where

$$\hat{\lambda}_{\mathrm{MSE}} = C_3(K) \left\{ \frac{\sigma_Q^4(\widehat{N})(b - a)}{\hat{\theta}_{22}^{.05}(\hat{g}_{\mathrm{MSE}})^2 n^2} \right\}^{1/9}.$$

(iii) The selected bandwidth is

$$\hat{h}_{\mathrm{DPI}} = C_1(K)\left\{\frac{\hat{\sigma}^2(\hat{\lambda}_{\mathrm{MSE}})(b-a)}{\hat{\theta}_{22}^{.05}(\hat{g}_{\mathrm{MSE}})n}\right\}^{1/5}.$$

Ruppert *et al.* (1995) demonstrated that this bandwidth selector performs well in practice. In terms of theoretical performance, they showed that $\hat{h}_{\mathrm{DPI}}/h_{\mathrm{MISE}} - 1 = O_P(n^{-2/7})$. This $O_P(n^{-2/7})$ relative rate is a very substantial improvement on cross-validation selectors, which have an $O_P(n^{-1/10})$ rate (Härdle *et al.*, 1988).

We now return to the problem of bandwidth selection for estimating curves from mixture data. Our method is based on that of Ruppert *et al* (1995).

The model for generating mixture data from curves was described in Section 5.3. There we stated that $X_1, \ldots, X_n$ were independent random variables with a common continuous distribution. In this problem it is convenient to consider that the X_i's, which have a corresponding value of I_i equal to 1, have a common density f_1 and the remaining X_i's have a common density f_2. Both f_1 and f_2 have their support confined to a compact interval $[a, b]$. For our method it greatly simplifies matters if we assume that we observe $I_1, \ldots, I_n$. This assumption reduces the curve estimation problem to the more tractable one of estimating two separate regression curves. This allows us to borrow ideas from existing bandwidth selectors for local linear estimators.

We define the conditional weighted mean integrated squared error of $\widehat{m}_1(\cdot\,; h_1)$ to be

$$\begin{aligned}&\mathrm{MISE}\left\{\widehat{m}_1(\cdot\,; h_1) \,|\, X_1, \ldots, X_n, I_1, \ldots, I_n\right\}\\ &\qquad = E\left[\int \left\{\widehat{m}_1(x; h) - m_1(x)\right\}^2 f_1(x)\, dx \,\middle|\, X_1, \ldots, X_n, I_1, \ldots, I_n\right],\end{aligned}$$

The asymptotic approximation of this MISE is

$$\begin{aligned}&\mathrm{MISE}\left\{\widehat{m}_1(\cdot\,; h_1) \,|\, X_1, \ldots, X_n\right\}\\ &\qquad = (n\pi)^{-1}h_1^{-1}R(K)(b-a)\sigma_1^2 + \frac{1}{4}h_1^4\mu_2(K)^2\int m_1''(x)^2 f_1(x)\, dx\\ &\qquad\qquad + o_p(n^{-1}h_1^{-1} + h_1^4).\end{aligned}$$

Thus,

$$h_{1,\mathrm{MISE}} \approx C_1(K)\left\{\frac{(b-a)\sigma_1^2}{\theta_{1,22}\, n\pi}\right\}^{1/5}.$$

Similarly we may obtain

$$h_{2,\mathrm{MISE}} \approx C_1(K)\left\{\frac{(b-a)\sigma_2^2}{\theta_{2,22}\, n(1-\pi)}\right\}^{1/5},$$

where $\theta_{l,rs} = \int m_l^{(r)}(x) m_l^{(s)}(x) f_l(x)\, dx$, for $l = 1, 2$, $r, s \geq 0$ and $r + s$ even.

If we had observed $I_1, \ldots, I_n$ then the natural kernel estimators of $\theta_{1,rs}$ and $\theta_{1,rs}$ would be

$$\hat{\theta}_{1,rs}(g) = (n\pi)^{-1} \sum_{i=1}^{n} I_i\, \widehat{m}_1^{(r)}(X_i; g)\, \widehat{m}_1^{(s)}(X_i; g)$$

and

$$\hat{\theta}_{2,rs}(g) = [n(1-\pi)]^{-1} \sum_{i=1}^{n} (1 - I_i)\, \widehat{m}_2^{(r)}(X_i; g)\, \widehat{m}_2^{(s)}(X_i; g).$$

The asymptotically MSE-optimal bandwidth g would be

$$g_{1,\mathrm{MSE}} = C_2(K)\left\{\frac{\sigma_1^2(b-a)}{|\hat{\theta}_{1,24}|\, n\pi}\right\}^{1/7}$$

and

$$g_{2,\mathrm{MSE}} = C_2(K)\left\{\frac{\sigma_2^2(b-a)}{|\hat{\theta}_{2,24}|\, n(1-\pi)}\right\}^{1/7}.$$

We do not observe the values of the I_i's but since we plan to use this bandwidth selector within the EM-style algorithm which we developed in Section 5.3, we shall have the z_i's, the expected values of the I_i's, conditional on our current curve estimates. The value of the z_i's can be calculated by equation (5.8). Replacing the I_i's with the z_i's, and allowing for the option of truncating the data near the boundaries, yields the new estimators

$$\hat{\theta}^{\alpha}_{1,rs}(g) = (n\pi)^{-1} \sum_{i=1}^{n} z_i\, \widehat{m}_1^{(r)}(X_i; g)\, \widehat{m}_1^{(s)}(X_i; g) 1_{\{(1-\alpha)a+\alpha b < X_i < \alpha a+(1-\alpha)b\}}$$

and

$$\hat{\theta}^{\alpha}_{2,rs}(g) = \{n(1-\pi)\}^{-1} \sum_{i=1}^{n} (1 - z_i)\, \widehat{m}_2^{(r)}(X_i; g)\, \widehat{m}_2^{(s)}(X_i; g) 1_{\{(1-\alpha)a+\alpha b < X_i < \alpha a+(1-\alpha)b\}}.$$

Similarly, the natural estimator for σ_1^2 is

$$\hat{\sigma}_1^2(\lambda_1) = \nu_1^{-1} \sum_{i=1}^{n} z_i \{Y_i - \widehat{m}_1(X_i; \lambda_1)\}^2, \tag{5.13}$$

with $\nu_1 = n\pi - 2\sum_i w_{1,ii} + \sum\sum_{ij} w_{1,ij}^2$ and

$$w_{1,ij} = \frac{\{\hat{s}_{21}(X_i;\lambda_1) - \hat{s}_{11}(X_i;\lambda_1)(X_i - x)\}\, z_i\, K\{(X_i - x)/\lambda_1\}}{\hat{s}_{21}(X_i;\lambda_1)\hat{s}_{01}(X_i;\lambda_1) - \hat{s}_{11}(X_i;\lambda_1)^2}.$$

The estimator for σ_2^2 is

$$\hat{\sigma}_2^2(\lambda_2) = \nu_2^{-1} \sum_{i=1}^{n} (1 - z_i)\, \{Y_i - \widehat{m}_2(X_i;\lambda_2)\}^2, \tag{5.14}$$

with $\nu_2 = n(1-\pi) - 2\sum_i w_{2,ii} + \sum\sum_{ij} w_{2,ij}^2$ and

$$w_{2,ij} = \frac{\{\hat{s}_{22}(X_i;\lambda_2) - \hat{s}_{12}(X_i;\lambda_2)(X_i - x)\}\, z_i\, K\{(X_i - x)/\lambda_2\}}{\hat{s}_{22}(X_i;\lambda_2)\hat{s}_{02}(X_i;\lambda_2) - \hat{s}_{12}(X_i;\lambda_2)^2}.$$

The $\hat{s}_{rl}$ are defined by equations (5.9) and (5.11), for $l = 1, 2$ respectively. The estimators of the optimal bandwidths $\lambda_{1,\mathrm{MSE}}$ and $\lambda_{2,\mathrm{MSE}}$ are given below, in the statement of the algorithm.

The blocked quartic estimator $\widehat{m}_1^Q(x; N_1)$ is obtained by fitting a *weighted* least-squares quartic fit to each of N blocks. The i'th weight for this fit is z_i. The blocked quartic estimator $\widehat{m}_2^Q(x; N_2)$ is obtained in the same fashion, except that here the i'th weight for the least-squares fit is $(1 - z_i)$. The blocked estimators for σ_1^2 and σ_2^2 are given by

$$\hat{\sigma}_{1,Q}^2(N_1) = (n\pi - 5N_1)^{-1} \sum_{i=1}^{n} z_i \{Y_i - \widehat{m}_1^Q(X_i; N_1)\}^2$$

and

$$\hat{\sigma}_{2,Q}^2(N_2) = \{n(1-\pi) - 5N_2\}^{-1} \sum_{i=1}^{n} (1 - z_i) \{Y_i - \widehat{m}_2^Q(X_i; N_2)\}^2.$$

Similarly,

$$\hat{\theta}_{1,rs}^Q(N_1) = (n\pi)^{-1} \sum_{i=1}^{n} z_i (\widehat{m}_1^Q)^{(r)}(X_i; N_1)(\widehat{m}_1^Q)^{(s)}(X_i; N_1)$$

and

$$\hat{\theta}_{2,rs}^Q(N_2) = \{n(1-\pi)\}^{-1} \sum_{i=1}^{n} (1 - z_i)(\widehat{m}_2^Q)^{(r)}(X_i; N_2)(\widehat{m}_2^Q)^{(s)}(X_i; N_2).$$

The number of blocks $\widehat{N}_1$ for the blocked quartic estimator $\widehat{m}_1^Q(x; N_1)$ is chosen to minimise

$$C_p(N_1) = \mathrm{RSS}_1(N_1)/\{\mathrm{RSS}_1(N_{\max})/(n\pi - 5N_{\max})\} - (n\pi - 10N_1),$$

where $\mathrm{RSS}_1(N_1)$ is the *weighted* residual sum of squares for $\widehat{m}_1^Q(x; N_1)$. Similarly, $\widehat{N}_2$ is chosen to minimise

$$C_p(N_2) = \mathrm{RSS}_2(N_2)/[\mathrm{RSS}_2(N_{\max})/\{n(1-\pi) - 5N_{\max}\}] - \{n(1-\pi) - 10N_2\},$$

where $\mathrm{RSS}_2(N_2)$ is the *weighted* residual sum of squares for $\widehat{m}_2^Q(x; N_2)$.

Now these estimators have been defined we can outline the algorithm for selecting the bandwidths $\hat{h}_{1,\mathrm{DPI}}$ and $\hat{h}_{2,\mathrm{DPI}}$ for the two curve estimators $\widehat{m}_1(\cdot\,; h_1)$ and $\widehat{m}_2(\cdot\,; h_2)$.

(i) Find $\hat{\theta}_{1,24}^Q(\widehat{N}_1), \hat{\theta}_{2,24}^Q(\widehat{N}_2), \hat{\sigma}_{1,Q}^2(\widehat{N}_1)$ and $\hat{\sigma}_{2,Q}^2(\widehat{N}_2)$ based on a blocked quartic fits with $\widehat{N}_1$ and $\widehat{N}_2$ chosen by Mallows' C_p.

(ii) Estimate $\theta_{1,22}$ using $\hat{\theta}_{1,22}^{.05}(\hat{g}_{1,\mathrm{MSE}})$, where

$$\hat{g}_{1,\mathrm{MSE}} = C_2(K)\left\{\frac{\sigma_{1,Q}^2(\widehat{N}_1)(b-a)}{|\hat{\theta}_{1,24}^Q(\widehat{N}_1)|\, n\pi}\right\}^{1/7}.$$

Similarly, estimate $\theta_{2,22}$ using $\hat{\theta}_{2,22}^{.05}(\hat{g}_{2,\mathrm{MSE}})$, where

$$\hat{g}_{2,\mathrm{MSE}} = C_2(K)\left\{\frac{\sigma_{2,Q}^2(\widehat{N}_2)(b-a)}{|\hat{\theta}_{2,24}^Q(\widehat{N}_2)|\, n(1-\pi)}\right\}^{1/7}.$$

Estimate σ_1^2 using $\hat{\sigma}_1^2(\hat{\lambda}_{1,\mathrm{MSE}})$, where

$$\hat{\lambda}_{1,\mathrm{MSE}} = C_3(K)\left\{\frac{\sigma_{1,Q}^4(\widehat{N}_1)(b-a)}{\hat{\theta}_{1,22}^{.05}(\hat{g}_{1,\mathrm{MSE}})^2(n\pi)^2}\right\}^{1/9}, \tag{5.15}$$

and estimate σ_2^2 using $\hat{\sigma}_2^2(\hat{\lambda}_{2,\mathrm{MSE}})$, where

$$\hat{\lambda}_{2,\mathrm{MSE}} = C_3(K)\left\{\frac{\sigma_{2,Q}^4(\widehat{N}_2)(b-a)}{\hat{\theta}_{2,22}^{.05}(\hat{g}_{2,\mathrm{MSE}})^2[n(1-\pi)]^2}\right\}^{1/9}. \tag{5.16}$$

(iii) The selected bandwidths are

$$\hat{h}_{1,\mathrm{DPI}} = C_1(K)\left\{\frac{\hat{\sigma}_1^2(\hat{\lambda}_{1,\mathrm{MSE}})(b-a)}{\hat{\theta}_{1,22}^{.05}(\hat{g}_{1,\mathrm{MSE}})\,n\pi}\right\}^{1/5} \tag{5.17}$$

and

$$\hat{h}_{2,\mathrm{DPI}} = C_1(K)\left\{\frac{\hat{\sigma}_2^2(\hat{\lambda}_{2,\mathrm{MSE}})(b-a)}{\hat{\theta}_{2,22}^{.05}(\hat{g}_{2,\mathrm{MSE}})\,n(1-\pi)}\right\}^{1/5}. \tag{5.18}$$

The derivation of this algorithm was based on the assumption that we observed the values $I_1, \ldots, I_n$. This was done to simplify what would otherwise be an extremely involved problem. The main consequence of this assumption is that the bandwidths $\hat{h}_{1,\mathrm{DPI}}$ and $\hat{h}_{2,\mathrm{DPI}}$ will, in general, undersmooth the data. By assuming that we have more information than we actually observe, we underestimate the variances of $\widehat{m}_1(\cdot\,; h_1)$ and $\widehat{m}_2(\cdot\,; h_2)$. This underestimation results in selection of bandwidths which are smaller than optimal. When the curves m_1 and m_2 are well-separated we do not lose much information by not observing which data points belong to which curves, and hence the amount of undersmoothing will be negligible. When the curves are not so well-separated the degree of undersmoothing will be greater.

A possible alternative to our approach would be a local likelihood-based method; see Fan and Gijbels (1996, Section 4.9). This approach would be likely to be extremely complicated, both theoretically and in practical implementation, since mixture likelihoods are notoriously difficult to work with, even in the simplest of situations; for example, see the discussion of singularities in the mixture likelihood function in Section 1.8. Another drawback of this approach is that since it is based entirely on a Normal mixture likelihood it would probably be more sensitive to departures from Normality than our technique. Our bandwidth selector seems to work very reasonably in practice.

5.5 Assessing the number of components in a mixture of smooth curves

We observed in Section 1.8 that the problem of determining the number of components in a mixture is a very difficult one, for which no clear statistical procedure has

been developed. If the components of a mixture are unimodal and sufficiently far apart then the number of modes in a distribution indicates the number of components in the mixture. In this section we shall assume that the component densities are unimodal and well enough separated for multimodality to be discernable. We shall use the calibrated bandwidth test, developed and studied in Chapters 2, 3 and 4, to assess multimodality.

First we consider the case of testing the null hypothesis that the independent data pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ are generated from a single smooth curve m, using the model

$$Y_i = m(X_i) + \epsilon_i,$$

against the alternative that these data are generated from more than one curve. We assume that the ϵ_i's are independent, and have zero mean and a common unimodal density. This setup implies that the errors are homoscedastic but adjustments can be made to deal with heteroscedasticity.

We propose the following method to test this null hypothesis.

(i) Fit a local linear estimator $\widehat{m}(\cdot\,; \hat{h}_{\mathrm{DPI}})$ to the data $(X_1, Y_1), \ldots, (X_n, Y_n)$. The bandwidth $\hat{h}_{\mathrm{DPI}}$ is chosen using the direct plug-in bandwidth selector of Ruppert *et al.* (1995), which was described in Section 5.4.

(ii) Calculate the residuals

$$r_i = Y_i - \widehat{m}(X_i; \hat{h}_{\mathrm{DPI}}).$$

(iii) Perform the calibrated bandwidth test for unimodality on the residuals. If we accept the null hypothesis that the residuals are unimodal then we accept the null hypothesis that the data are generated from a single smooth curve.

In step (iii) we use the form of the bandwidth test that has been calibrated using a sample size-based method. This method was called form (e) in Chapters 3 and 4. Departures from the assumption of homoscedastic errors may be accommodated by dividing the residuals r_i by $\hat{\sigma}(X_i)$, where $\hat{\sigma}^2(\cdot)$ is some estimate of the variance function.

A similar testing procedure may be developed for testing the null hypothesis that the data pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ are generated from a mixture of two smooth

curves, $m_1(\cdot)$ and $m_2(\cdot)$, versus the alternative that these data are generated from three or more curves. See the beginning of Section 5.3 for the exact description of the model under the null. Our approach is based on partitioning the data pairs into those that belong to the curve m_1 and those that belong to m_2. This is done by estimating m_1 and m_2 and calculating the z_i's using equation (5.8). Remember that

$$z_i = P\{I_i = 1 \,|\, (X_i, Y_i), \widehat{m}_1(\cdot\,; \hat{h}_{\mathrm{DPI}}), \widehat{m}_2(\cdot\,; \hat{h}_{\mathrm{DPI}}), \hat{\pi}, \hat{\sigma}_1, \hat{\sigma}_2\},$$

that is, the probability that the pair (X_i, Y_i) belongs to the curve m_1, conditional on the data and our estimates. If $z_i > 0.5$ then we attribute the pair to the curve m_1, otherwise we attribute it to the curve m_2. Then we can apply the calibrated bandwidth test for unimodality to each of the two sets of residuals. The exact steps are:

(i) Use the algorithm of Section 5.3 to fit local linear estimators $\widehat{m}_1(\cdot\,; \hat{h}_{1,\mathrm{DPI}})$ and $\widehat{m}_2(\cdot\,; \hat{h}_{2,\mathrm{DPI}})$ to the data and compute the values of the z_i's.

(ii) For the pairs (X_i, Y_i) whose corresponding $z_i > 0.5$, calculate the residuals

$$r_{1i} = Y_i - \widehat{m}_1(X_i; \hat{h}_{1,\mathrm{DPI}}),$$

and call this set of residuals $\mathcal{X}_1$. For the remaining data calculate the residuals

$$r_{2i} = Y_i - \widehat{m}_2(X_i; \hat{h}_{2,\mathrm{DPI}}),$$

and call this set of residuals $\mathcal{X}_2$.

(iii) Perform the calibrated bandwidth test for unimodality on each of $\mathcal{X}_1$ and $\mathcal{X}_2$ separately. If we accept the null hypothesis of unimodality for both samples, $\mathcal{X}_1$ and $\mathcal{X}_2$, then we accept the null hypothesis that the data are generated from a mixture of two smooth curves.

This approach may be naturally extended to testing for the presence of three or more curves. We study the practical effectiveness of these tests in the simulation study in Section 5.6.

5.6 Practical performance

We conducted a simulation study to evaluate the performance of the algorithm for estimating the mean curves from mixture data and to examine the effectiveness

of the calibrated bandwidth test for determining the number of components in a mixture of smooth curves.

Throughout this study the X_i's were generated from the Uniform distribution on $[0,1]$. The binary random variables $I_1, \ldots, I_n$ were assigned the value zero with probability π, and the value 1 with probability $1-\pi$. The Y_i's were generated by

$$Y_i = \begin{cases} m_1(X_i) + \sigma_1 \epsilon_i & \text{if } I_i = 1 \\ m_2(X_i) + \sigma_2 \epsilon_i & \text{if } I_i = 0, \end{cases}$$

where the ϵ_i's were random variates from a standard Normal distribution. For all our examples, the standard deviations σ_1 and σ_2 were both taken to be 0.1, the mixing proportion π was set at 0.5 and the curve m_1 was given by

$$m_1(x) = 3\exp\{-x^2/(0.3)^2\} + 3\exp\{-(x-1)^2/(0.7)^2\}.$$

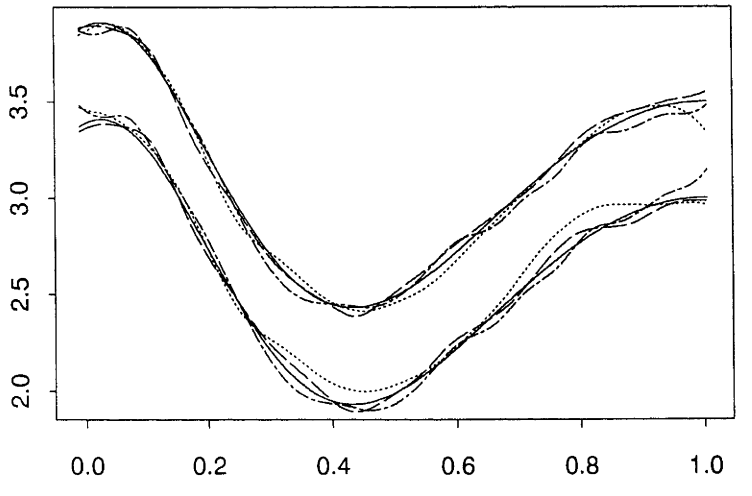
Three different examples were considered for the second curve m_2. These were:

Example 1: $m_2(x) = 3\exp\{-x^2/(0.3)^2\} + 3\exp\{-(x-1)^2/(0.7)^2\} + 0.5$,
Example 2: $m_2(x) = 3\exp\{-(x-0.2)^2/(0.3)^2\} + 3\exp\{-(x-1.2)^2/(0.7)^2\}$,
Example 3: $m_2(x) = 6\exp\{-(x-0.2)^2/(0.3)^2\} + 6\exp\{-(x-1.2)^2/(0.7)^2\} - 2.7$.

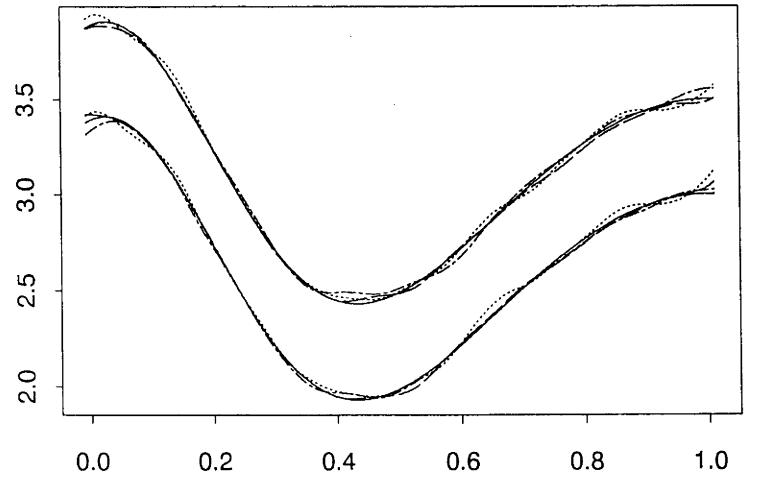
In Example 1 the two curves were parallel, in Example 2 m_1 and m_2 were identical up to a small phase change, and Example 3 was the same as Example 2 except that the amplitude of m_2 was doubled. Sample sizes of $n = 200$ and $n = 1000$ were used. The number of replications in the simulation was 100. The I_i's were taken as the starting values for the z_i's in the algorithm for estimating the curves.

The results of the simulations are summarised in Figures 5.1–5.3 and Table 5.1. In each figure, plot (a) shows the true curves (solid lines) and three estimates for a sample size of $n = 200$ and plot (b) displays the same for the larger sample size of $n = 1000$. The three estimated curves were chosen to give an indication of the range of the quality of the curve estimates. We measured the accuracy of the fit of a pair of estimates, $\widehat{m}_1$ and $\widehat{m}_2$ by a combined integrated squared error (ISE), given by

$$\begin{aligned} \text{ISE}\,(\widehat{m}_1, \widehat{m}_2) \;&=\; \pi \int \{m_1(x) - \widehat{m}_1(x)\}^2 f(x)\,dx \\ &\quad + (1-\pi) \int \{m_2(x) - \widehat{m}_2(x)\}^2 f(x)\,dx\,. \end{aligned}$$

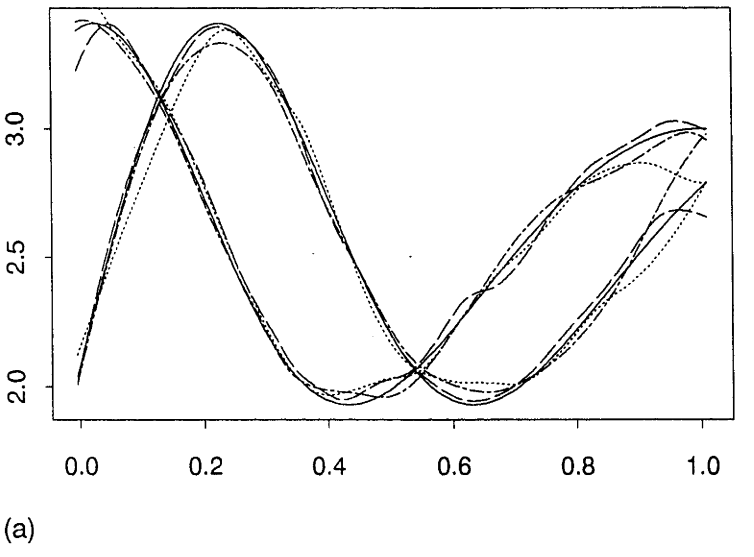


(a)

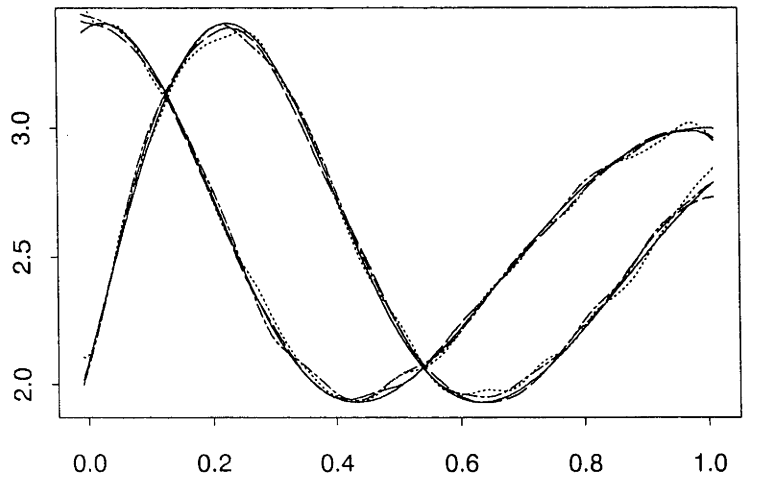


(b)

Figure 5.1: *Graphical summary of the simulation results for Example 1 for sample sizes (a)* $n = 200$ *and (b)* $n = 1000$. *The four pairs of curves on each plot represent the true pair of curves (solid lines) and the 10th best (dashed lines), 50th best (dot-dashed lines) and 90th best (dotted lines) fits out of the 100 computed estimates.*



(a)



(b)

Figure 5.2: *Graphical summary of the simulation results for Example 2 for sample sizes (a)* $n = 200$ *and (b)* $n = 1000$. *The four pairs of curves on each plot represent the true pair of curves (solid lines) and the 10th best (dashed lines), 50th best (dot-dashed lines) and 90th best (dotted lines) fits out of the 100 computed estimates.*

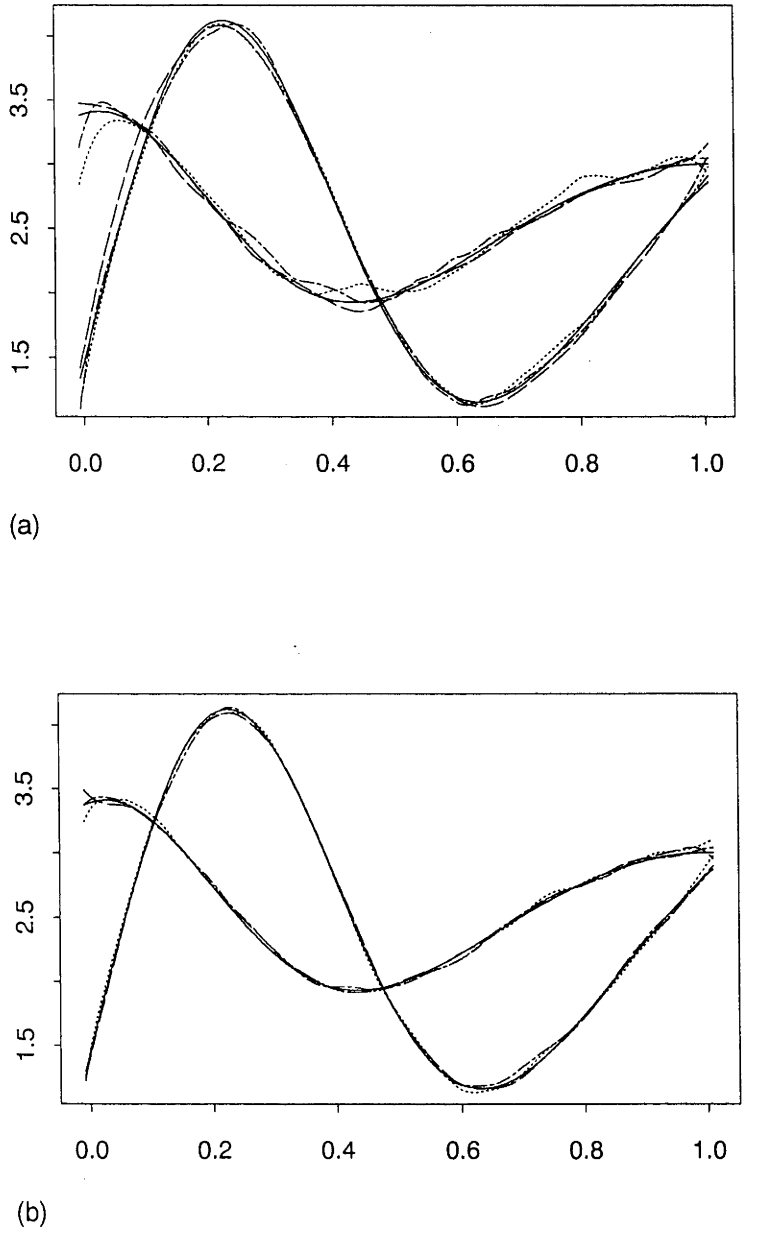


Figure 5.3: *Graphical summary of the simulation results for Example 3 for sample sizes (a)* $n = 200$ *and (b)* $n = 1000$. *The four pairs of curves on each plot represent the true pair of curves (solid lines) and the 10th best (dashed lines), 50th best (dot-dashed lines) and 90th best (dotted lines) fits out of the 100 computed estimates.*

Table 5.1: *Means and standard deviations of parameter estimates for the three examples.*

	Sample size	Estimator	True value	Mean	Std dev.
Example 1	$n = 200$	$\hat{\pi}$	0.5	0.4943	0.04157
		$\hat{\sigma}_1^2$	0.01	0.01011	0.002062
		$\hat{\sigma}_2^2$	0.01	0.01019	0.001823
	$n = 1000$	$\hat{\pi}$	0.5	0.4993	0.01607
		$\hat{\sigma}_1^2$	0.01	0.01007	0.0008121
		$\hat{\sigma}_2^2$	0.01	0.009964	0.0007432
Example 2	$n = 200$	$\hat{\pi}$	0.5	0.5031	0.04232
		$\hat{\sigma}_1^2$	0.01	0.01045	0.001996
		$\hat{\sigma}_2^2$	0.01	0.01047	0.002091
	$n = 1000$	$\hat{\pi}$	0.5	0.5003	0.01685
		$\hat{\sigma}_1^2$	0.01	0.01022	0.0007057
		$\hat{\sigma}_2^2$	0.01	0.01006	0.0007907
Example 3	$n = 200$	$\hat{\pi}$	0.5	0.4984	0.03780
		$\hat{\sigma}_1^2$	0.01	0.01030	0.001725
		$\hat{\sigma}_2^2$	0.01	0.01034	0.001725
	$n = 1000$	$\hat{\pi}$	0.5	0.4993	0.01703
		$\hat{\sigma}_1^2$	0.01	0.01013	0.0006995
		$\hat{\sigma}_2^2$	0.01	0.01014	0.0007773

The three pairs of estimated curves on the plots represent the 10th best (dashed lines), 50th best (dot-dashed lines) and 90th best (dotted lines) fits, in terms of ISE, out of the 100 computed estimates.

As predicted in Section 5.4, our bandwidth selector tended to undersmooth the data a little, especially in the $n = 200$ cases, but its performance was quite reasonable. As with standard curve estimation procedures, out estimates performed worst near turning points and the boundaries. In addition to these troublesome points, our estimates experienced difficulties at points where the two curves intersected; at these points the closeness of the components of the mixture makes estimation difficult. Table 5.1 lists the means and standard deviations of the estimates of the

Table 5.2: *The power for testing the null hypothesis that the data are generated from a single smooth curve, versus the alternative that they are generated from a mixture of two or more curves.*

		Level of the test			
	Sample size	0.01	0.05	0.1	0.2
Example 1	$n = 200$	1.00	1.00	1.00	1.00
	$n = 1000$	1.00	1.00	1.00	1.00
Example 2	$n = 200$	0.37	0.54	0.63	0.74
	$n = 1000$	0.95	0.99	1.00	1.00
Example 3	$n = 200$	0.66	0.82	0.90	0.93
	$n = 1000$	1.00	1.00	1.00	1.00

parameters π, σ_1^2 and σ_2^2. Taking sample size into account, our methods appear to perform very well overall for the three examples.

In our simulation study we also examined the usefulness of the calibrated bandwidth test, which was developed and studied in Chapters 2, 3 and 4, as a tool for determining the number of components in a mixture of smooth curves. This approach is based on applying the bandwidth test to residuals and is outlined in Section 5.5 .

We first tested the null hypothesis that the data were generated from a single smooth curve. For each example we performed the test on each of the 100 simulated data sets. The power for the test at each of the levels $\alpha = 0.01, 0.05, 0.1$ and 0.2 was estimated by the proportion of times the null hypothesis was rejected. These power estimates are recorded in Table 5.2 .

For Example 1 the null hypothesis was correctly rejected for all 100 samples for both sample sizes $n = 200$ and $n = 1000$. In this example the two curves are parallel and separated by a distance of 0.5, which is five times the component standard deviations of 0.1 . Since the two curves are very well separated, the test has no trouble detecting the presence of more than one curve.

In Examples 2 and 3 the two curves intersect twice, so there are regions where the two curves are not well separated. For the larger sample size, $n = 1000$, the test had full or almost full power. For the smaller sample size, $n = 200$, the test had

some power – 63% at the 0.1 level – for Example 2, and quite good power – 90% at the 0.1 level – for Example 3.

Next we tested the null hypothesis that the data were generated from a mixture of two curves. Since this null hypothesis was true we wished to study the level accuracy of the test in this case. We calculated the residuals, partitioned them and performed the calibrated bandwidth test as described in Section 5.5 . The level accuracy for each of the samples is summarised graphically in Figures 5.4–5.6. In all cases the level accuracy is very good. If the calibrated version of the test had not been used, then the level accuracy would have been very poor and, more importantly, the test would have had far less power in the case of testing for the null hypothesis that the data is generated from a single smooth curve. Our examples indicate that the calibrated bandwidth test can be an effective tool for determining the number of smooth curves from which data are generated.

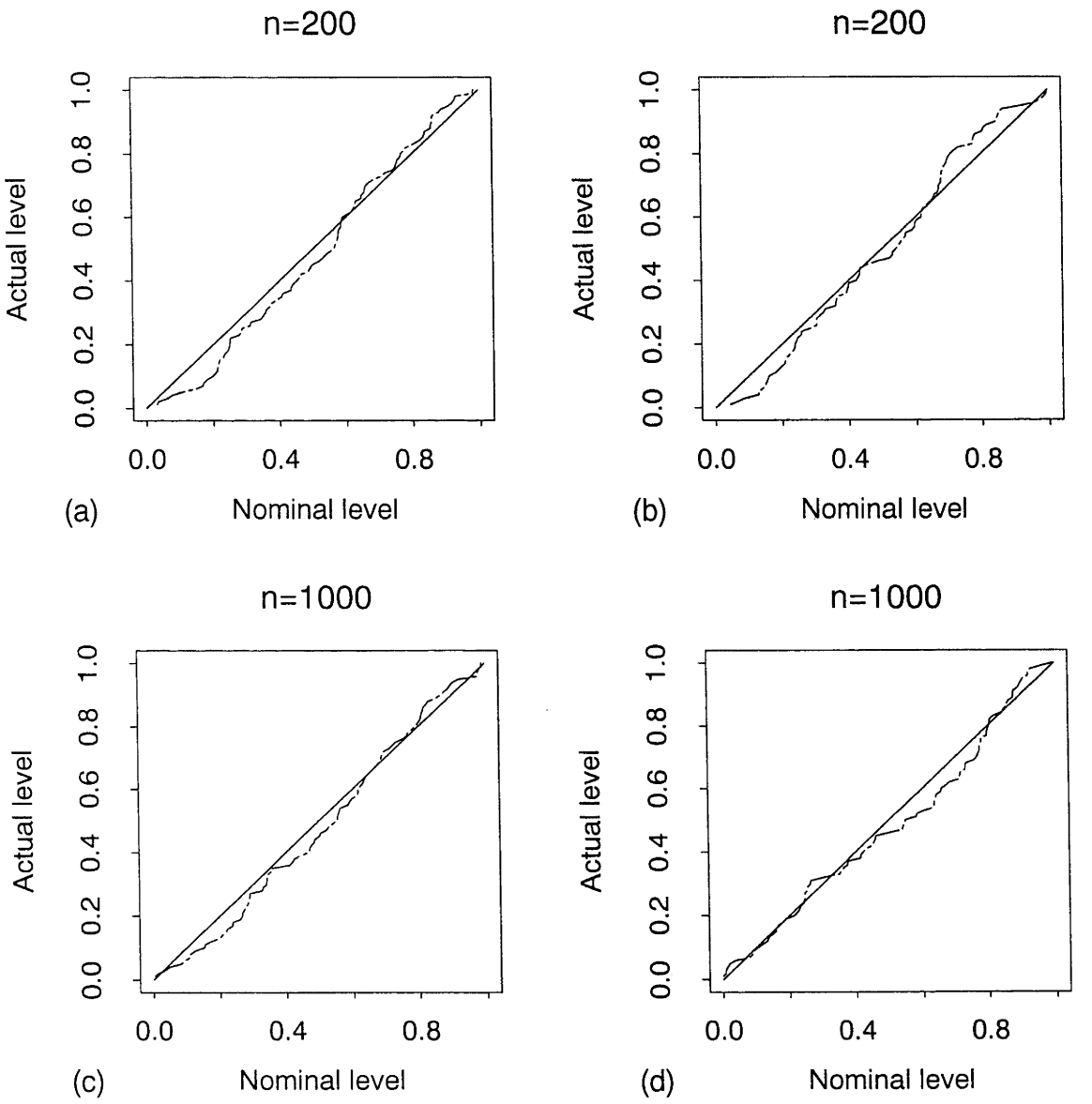


Figure 5.4: *Level accuracy for testing the null hypothesis that the residuals for Example 1 are unimodal. Panel (a) displays the results for the residuals associated with the curve m_1, for a sample of size $n = 200$. Panel (b) shows the same for the residuals associated with the curve m_2. Panels (c) and (d) plot the same results as panels (a) and (b) but for the larger sample size of $n = 1000$.*

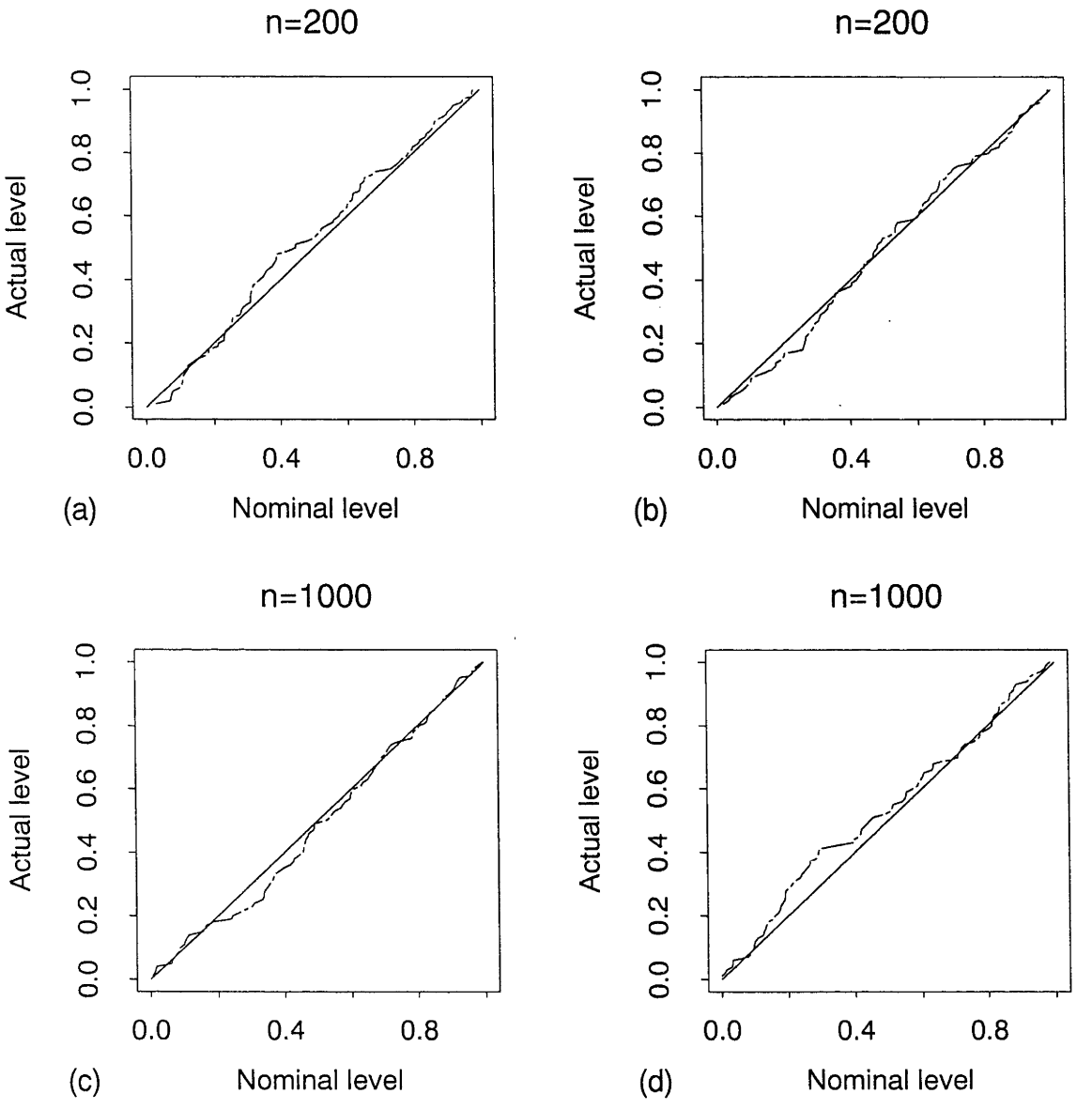


Figure 5.5: *Level accuracy for testing the null hypothesis that the residuals for Example 2 are unimodal. Panel (a) displays the results for the residuals associated with the curve* m_1*, for a sample of size* $n = 200$*. Panel (b) shows the same for the residuals associated with the curve* m_2*. Panels (c) and (d) plot the same results as panels (a) and (b) but for the larger sample size of* $n = 1000$*.*

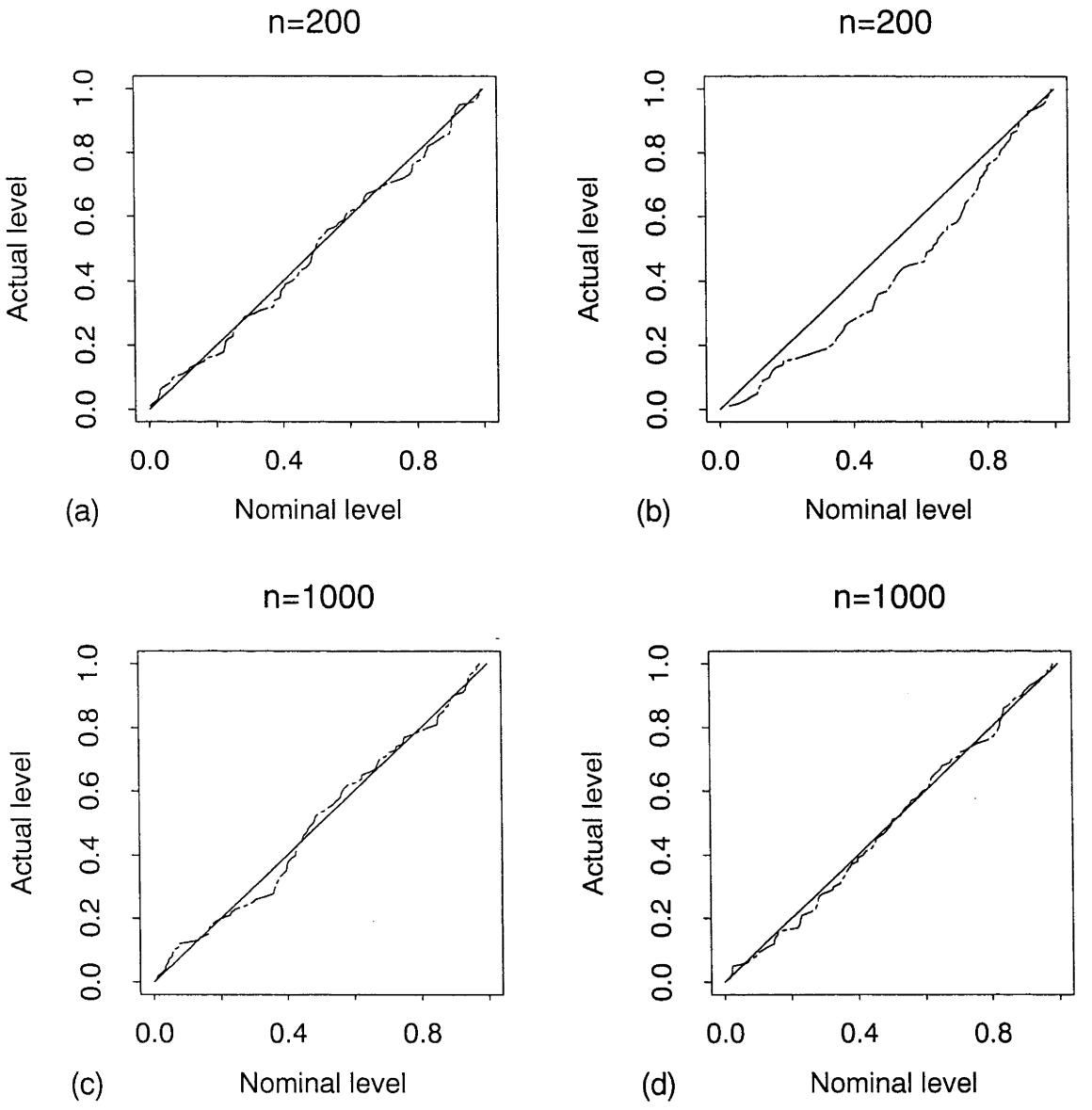


Figure 5.6: *Level accuracy for testing the null hypothesis that the residuals for Example 3 are unimodal. Panel (a) displays the results for the residuals associated with the curve m_1, for a sample of size $n = 200$. Panel (b) shows the same for the residuals associated with the curve m_2. Panels (c) and (d) plot the same results as panels (a) and (b) but for the larger sample size of $n = 1000$.*

References

Cheng, M.-Y. and Hall, P. (1997). Calibrating the excess mass test of modality. *Research Report SRR 002-97.* Centre for Mathematics and its Applications, Australian National University, Canberra.

Cheng, M.-Y. and Hall, P. (1998). Calibrating the excess mass and dip tests of modality. *J. Roy. Statist. Soc. Ser. B* **60**, 579–589.

Choi, K. and Bulgren, W.B. (1968). An estimation procedure for mixtures of distributions. *J. Roy. Statist. Soc. Ser. B* **30**, 444–460.

Cox, D.R. (1966). Notes on the analysis of mixed frequency distributions. *Brit. J. Math. Statist. Psychol.* **19**, 39–47.

Cox, D.R. and Hinkley, D.V. (1974). *Theoretical Statistics.* Chapman and Hall, London.

Cuevas, A. and Gonzalez-Manteiga, W. (1991). Data-driven smoothing based on convexity properties. In *Nonparametric Functional Estimation and Related Topics,* G. Roussas (ed.), pp. 225–240. Kluwer, Dordrecht.

Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39**, 1–38.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7**, 1–26.

Efron, B. and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap.* Chapman and Hall, New York.

Escobar, M.D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* **90**, 577–588.

Everitt, B.S. and Hand, D.J. (1981). *Finite Mixture Distributions.* Chapman and Hall, London.

Fan, J. (1992). Design-adaptive nonparametric regression. *J. Amer. Statist. Assoc.* **87**, 998–1004.

Fan, J. (1993). Local linear regression smoothers and their minimax efficiency. *Ann. Statist.* **21**, 196–216.

Fan, J. and Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: Variable bandwidth and spatial adaption. *J. Roy. Statist. Soc. Ser. B* **57**, 371–394.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications.* Chapman and Hall, London.

Fan, J., Gijbels, I., Hu, T.-C. and Huang, L.-S. (1996). A study of variable bandwidth selection for local polynomial regression. *Statist. Sinica* **6**, 113–127.

Fan, J. and Marron, J.S. (1994). Fast implementations of nonparametric curve estimators. *J. Comp. Graph. Statist.* **3**, 35–56.

Fisher, N.I., Mammen, E. and Marron, J.S. (1994). Testing for modality. *Comput. Statist. and Data Analysis* **18**, 499–512.

Feller, W. (1966). *An Introduction to Probability Theory and its Applications, Volume II.* Wiley, New York.

Galambos, J. (1978). *The Asymptotic Theory of Extreme Order Statistics.* Wiley, New York.

Garsia, A.M. (1970). Continuity properties of a Gaussian process with multidimensional time parameter. *Proc. Sixth Berkeley Symp. Math. Statist. Probab.* **2**, 369–374. University of California Press.

Good, I.J. and Gaskins, R.A. (1971). Nonparametric roughness penalties for probability densities. *Biometrika* **58**, 255–277.

Good, I.J. and Gaskins, R.A. (1980). Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data. *J. Amer. Statist. Assoc.* **75**, 42–73.

Hall, P. (1986). On the bootstrap and confidence intervals. *Ann. Statist.* **14**, 1431–1452.

Hall, P. (1987). On the bootstrap and likelihood-based confidence intervals. *Biometrika* **74**, 481–493.

Hall, P. and Wand, M.P. (1996). On the accuracy of binned kernel density estimates. *J. Multiv. Anal.* **56**, 165–184.

Hall, P. and Wood, A.T.A. (1996). Approximations to distributions of statistics used for testing hypotheses about the number of modes of a population. *J. Statist. Planning and Inference* **55**, 299–317.

Härdle, W., Hall, P. and Marron, J.S. (1988). How far are automatically chosen regression smoothing parameters from their optimum? *J. Amer. Statist. Assoc.* **83**, 86–95.

Härdle, W. and Marron, J.S. (1995). Fast and simple scatterplot smoothing. *Comput. Statist. and Data Analysis* **20**, 1–17.

Hart, J.D. (1984). On the modal resolution of kernel density estimators. *Statist. Prob. Letters* **2**, 363–369.

Hartigan, J.A. (1981). Consistency of single linkage for high-density clusters. *J. Amer. Statist. Assoc.* **76**, 388–394.

Hartigan, J.A. (1985). A failure of likelihood asymptotics for normal mixtures. In *Proc. Berkeley Conf. in Honor of Jerzy Neyman and Jack Kiefer,* L.M. LeCam and R.A. Olshen (eds.), vol. II, pp. 807–810. Wadsworth and Brooks, Pacific Grove.

Hartigan, J.A. (1987). Estimation of a convex density contour in two dimensions. *J. Amer. Statist. Assoc.* **82**, 267–270.

Hartigan, J.A. (1988). The span test for unimodality. In *Classification and Related Methods of Data Analysis,* H. H. Bock (ed.), pp. 229–236. North-Holland, Amsterdam.

Hartigan, J.A. and Hartigan, P.M. (1985). The DIP test of unimodality. *Ann. Statist.* **13**, 70–84.

Hartigan, J. A. and Mohanty, S. (1992). The RUNT test for multimodality. *Journal of Classification* **9**, 63–70.

Hartigan, P.M. (1985). Algorithm AS 217. Computation of the DIP statistic for unimodality. *Appl. Statist.* **34**, 320–325.

Hathaway, R.J. (1985). A constrained formulation of maximum-likelihood estimation for normal mixture distributions. *Ann. Statist.* **13**, 795-800.

Hathaway, R.J. (1986). A constrained EM algorithm for univariate normal mixtures. *J. Statist. Comput. Simul.* **23**, 211–230.

Izenman, A.J. and Sommer, C. (1988). Philatelic mixtures and multimodal densities. *J. Amer. Statist. Assoc.* **83**, 941–953.

Johnson, N.L. and Kotz, S. (1970). *Continous Univariate Distributions,* Vol. 1. Houghton Mifflin, Boston.

Jones, M.C. (1989). Discretized and interpolated kernel density estimates. *J. Amer. Statist. Assoc.* **84**, 733–741.

Jones, M.C. (1991). On correcting for variance inflation in kernel density estimation. *Comput. Statist. and Data Analysis* **11**, 3–15.

Jones, M.C. and Lotwick, H.W. (1984). A remark on Algorithm AS 176. Kernel density estimation using the fast Fourier transform. Remark AS R50. *Appl. Statist.* **33**, 120–122.

Jones, M.C., Marron, J.S. and Sheather, S.J. (1996). A brief survey of bandwidth selection for density estimation. *J. Amer. Statist. Assoc.* **91**, 401–407.

Karlin, S. (1968). *Total Positivity.* Stanford University Press, Stanford.

Komlós, J., Majór, P. and Tusnády, G. (1975). An approximation of partial sums of independent RV's, and the sample DF. I. *Z. Wahrscheinlichkeitstheorie verw. Gebiete* **32**, 111–131.

Loh, W.-Y. (1987). Calibrating confidence intervals. *J. Amer. Statist. Assoc.* **82**, 155–162.

Loh, W.-Y. (1991). Bootstrap calibration for confidence interval construction and selection. *Statist. Sinica* **1**, 479–495.

Mallows, C.L. (1973). Some comments on C_p. *Technometrics* **15**, 661–675.

Mammen, E., Marron, J.S. and Fisher, N.I. (1992). Some asymptotics for multimodality tests based on kernel density estimates. *Probab. Theory Related Fields* **91**, 115–132.

Marron, J.S. and Nolan, D. (1989). Canonical kernels for density estimation. *Statist. Prob. Letters* **7**, 195–199.

McLachlan, G.J. (1987). On bootstrapping the the likelihood ratio test statistic for the number of components in a normal mixture. *Appl. Statist.* **36**, 318–324.

McLachlan, G.J. and Basford, K.E. (1988). Mixture Models: Inference and Applications to Clustering. Marcel Dekker, New York.

Minnotte, M.C. (1997). Nonparametric testing of the existence of modes. *Ann. Statist.* **25**, 1646–1660.

Minnotte, M.C. and Scott, D.W. (1993). The mode tree: a tool for visualization of nonparametric density features. *J. Comput. Graph. Statist.* **2**, 51–68.

Müller, D.W. and Sawitzki, G. (1987). Using excess mass estimates to investigate the modality of a distribution. *Sonderforschungsbereich 123.* Univerität Heidelberg, Heidelberg.

Müller, D.W. and Sawitzki, G. (1991). Excess mass estimates and tests for multimodality. *J. Amer. Statist. Assoc.* **86**, 738–746.

Parzen, E. (1979). Nonparametric statistical data modeling. *J. Amer. Statist. Assoc.* **74**, 105–131.

Pearson, K. (1894). Contribution to the mathematical theory of evolution. *Phil. Trans. Roy. Soc. A* **185**, 71–110.

Polonik, W. (1995a). Measuring mass concentrations and estimating density contour clusters – an excess mass approach. *Ann. Statist.* **23**, 855–881.

Polonik, W. (1995b). Density estimation under qualitative assumptions in higher dimensions. *J. Multiv. Anal.* **55**, 61–81.

Roeder, K. (1990). Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *J. Amer. Statist. Assoc.* **85**, 617–624.

Roeder, K. (1994). A graphical technique for determining the number of components in a mixture of normals. *J. Amer. Statist. Assoc.* **89**, 487–495.

Rozál, G.P.M. and Hartigan, J.A. (1994). The MAP test for multimodality. *J. Classification* **11**, 5–36.

Ruppert, D., Sheather, S.J. and Wand, M.P. (1995). An effective bandwidth selector for local least squares regression. *J. Amer. Statist. Assoc.* **90**, 1257–1270.

Schoenberg, I.J. (1950). On Pólya frequency functions, II: Variation diminishing integral operators of the convolution type. *Acta Scientiarum Mathematicarum*

12B, 97–106.

Scott, D.W. (1980). Comment on a paper by Good and Gaskins. *J. Amer. Statist. Assoc.* **75**, 61–62.

Scott, D.W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization.* Wiley, New York.

Scott, D.W., Gotto, A.M., Cole, J.S. and Gorry, G.A. (1978). Plasma lipids as collateral risk factors in coronary artery disease – a study of 371 males with chest pain. *J. Chronic Diseases* **31**, 337–345.

Simonoff, J.S. (1996). *Smoothing Methods in Statistics.* Springer, New York.

Silverman, B.W. (1978). Weak and strong uniform consistency of the kernel estimate of a density function and its derivatives. *Ann. Statist.* **6**, 177–184.

Silverman, B.W. (1981). Using kernel density estimates to investigate multimodality. *J. Roy. Statist. Soc. Ser. B* **43**, 97–99.

Silverman, B.W. (1982). Algorithm AS176. Kernel density estimation using the fast Fourier transform. *Appl. Statist.* **31**, 93–97.

Silverman, B.W. (1983). Some properties of a test for multimodality based on kernel density estimates. In *Probability, Statistics and Analysis,* J.F.C. Kingman and G.E.H. Reuter (eds.), pp. 248–259. Cambridge University Press, Cambridge.

Silverman, B.W. (1984). Spline smoothing: the equivalent variable kernel method. . *Ann. Statist.* **12**, 898–916.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

Silverman, B.W., Jones, M. C., Wilson, J. D. and Nychka, D. W. (1990). A smoothed EM approach to indirect estimation problems, with particular reference to stereology and emission tomography. *J. Roy. Statist. Soc. Ser. B* **52**, 271–324.

Sommer, C.J. and McNamara, J.N. (1987). Power considerations for the dip test of unimodality. *Comp. Sci. Statist.* **19**, 552-556.

Tapia, R.A. and Thompson, J.R. (1978). *Nonparametric Probability Density Estimation.* Johns Hopkins University Press, Baltimore.

Tibshirani, R. and Hastie, T.J. (1987). Local likelihood estimation. *J. Amer.*

Statist. Assoc. **82**, 559–567.

Titterington, D.M., Smith, A.F.M. and Makov, U.E. (1985). *Statistical Analysis of Finite Mixture Distributions.* Wiley, Chichester.

Wand, M.P. and Jones, M.C. (1995). *Kernel Smoothing.* Chapman and Hall, London.

Wu, C.F.J. (1983). On the convergence of the EM algorithm. *Ann. Statist.* **11**, 95–103.